

## Biogeography in Conservation: Tools to Explore the Past and Future of Species in a Changing World

Author(s): Carlos Alberto Arnillas, Sandy M. Smith, Felicity J. N., and Adam Martin

Source: *Lessons in Conservation*, Vol. 9, Issue 1, pp. 55-94

Published by: Network of Conservation Educators and Practitioners, Center for Biodiversity and Conservation, American Museum of Natural History

Stable URL: [ncep.amnh.org/linc](http://ncep.amnh.org/linc)

---

This article is featured in *Lessons in Conservation*, the official journal of the Network of Conservation Educators and Practitioners (NCEP). NCEP is a collaborative project of the American Museum of Natural History's Center for Biodiversity and Conservation (CBC) and a number of institutions and individuals around the world. *Lessons in Conservation* is designed to introduce NCEP teaching and learning resources (or "modules") to a broad audience. NCEP modules are designed for undergraduate and professional level education. These modules—and many more on a variety of conservation topics—are available for free download at our website, [ncep.amnh.org](http://ncep.amnh.org).



---

To learn more about NCEP, visit our website: [ncep.amnh.org](http://ncep.amnh.org).

All reproduction or distribution must provide full citation of the original work and provide a copyright notice as follows:

“Copyright 2019, by the authors of the material and the Center for Biodiversity and Conservation of the American Museum of Natural History. All rights reserved.”

Illustrations obtained from the American Museum of Natural History's library:



# Biogeography in Conservation: Tools to Explore the Past and Future of Species in a Changing World

Carlos Alberto Arnillas<sup>i</sup>, Sandy M. Smith<sup>ii</sup>, Felicity J. Ni<sup>iii</sup>, and Adam Martin<sup>iv</sup>

<sup>i</sup>University of Toronto Scarborough, Department of Physical and Environmental Sciences, Toronto, Ontario, CA; <sup>ii</sup>University of Toronto, Faculty of Forestry, Toronto, Ontario, CA; <sup>iii</sup>University of Toronto, School of the Environment, Toronto, Ontario, CA; <sup>iv</sup>University of Toronto Scarborough, Centre for Critical Development Studies, Toronto, Ontario, CA

## ABSTRACT

Humans have now altered essentially every natural ecosystem in the world, and among the numerous consequences of anthropogenic global change, many of the Earth's species are currently living under drastically different environmental and ecological conditions. On one hand, many species that once thrived in the wild are now threatened by extinction, while at the same time, species that were historically benign are becoming invasive in different parts of the world. To address this major challenge, it is critical that conservation practitioners understand the multiple short- and long-term climatological, geological, and evolutionary mechanisms that have resulted in the current distribution of species; understanding how these same mechanisms interact is also key in predicting species distributions—and possible extinctions—into the future. Using the Global Biodiversity Information Facility (GBIF), an open-access worldwide database of species occurrences, this research project exercise is designed to guide teams of students through the process of: a) identifying and researching characteristics relevant to understanding species distribution (e.g., age of the group, habitat requirements, dispersal capabilities); b) representing the present and historic species distribution; c) critically assessing the quality and amount of information available; d) using that information to understand species history and potential future challenges the species may either face or impose on the ecosystems; and e) sharing the results with peers and learning from that experience.

## LEARNING OBJECTIVES

After completing this research project exercise, students will be able to:

1. Understand the methods used to study both current and historical species distributions;
2. Use open-access biodiversity databases to record, map, and model the distribution of selected taxa;
3. Use basic database management tools to analyze, interpret, and communicate scientific data on species distributions graphically, orally, and in written reports;
4. Interpret and apply biodiversity data in the context of informing biodiversity conservation planning and policy; and finally,
5. Critically evaluate the potential use and limitations of open-access biodiversity databases, in understanding species historical and current distributions.

## BACKGROUND

### Biodiversity Conservation in the Anthropocene

The fate of animals, plants, and arguably entire ecosystems or biomes, is in the midst of unprecedented change as a result of human action (Ladle and Whittaker 2011, Pachauri et al. 2014). Currently, scientists, environmentalists, policy-makers, and citizens recognize that the structure, function, and composition of many habitats worldwide have been completely destroyed or severely damaged by human activity (Richardson and Whittaker 2010). Of the numerous impacts humans have had on the natural world, among the most prominent are both deliberate and accidental changes in the composition of species across all taxonomic kingdoms.

At the same time, anthropogenic climate change now threatens to completely redraw the geographic map of life on this planet. Increasingly, scientific evidence suggests that these massive, global anthropogenic changes in the composition of ecosystems are unique on geological timescales, which is one of the reasons our current geological era—the “Anthropocene” or the “Era of the Human”—is becoming more widely acknowledged as a distinct period of Earth's history (Crutzen 2002).

While some species are able to thrive in the Anthropocene, evidence suggests that the vast majority of species face a higher degree of susceptibility to anthropogenic changes, which in turn results



in: a) reduced local or global population numbers; b) increased rates of local or global extinction; and ultimately, c) increased likelihood of species compositional change in essentially all ecosystems globally (McKinney and Lockwood 1999). From a strictly human perspective, the consequences of such changes for ecosystems services that people directly rely on, such as clean air or water provisioning, are difficult to predict (Millennium Ecosystem Assessment 2005).

### **Integrating Conservation Biology and Biogeography**

Conservation biology—a term first coined as recently as 1978 (Douglas 1978)—is an applied field of scientific study that focuses on understanding spatial and temporal patterns in the Earth’s biodiversity, with the aim of managing species, their habitats, and ecosystems, in order to prevent their extinction (Soulé 1985, Soulé 1986). To predict the fate of key species and ecosystems, and ultimately prescribe solutions, conservation biology draws on tools and methods employed in many fields of study including fundamental taxonomy-based sciences, such as botany and zoology, as well as more quantitative predictive sciences, such as population and community ecology. Because potential solutions to prevent biodiversity loss must be implemented by people at different scales (e.g., local, regional, global), conservation biology is structured as an interdisciplinary and active research field, with critical participation from biologists, social scientists, economists, policy-makers, and engaged citizens.

Complementary to conservation biology, is biogeography: the field of study focused on understanding the processes that give rise to the spatial distribution of species. While biogeography and conservation biology undoubtedly share certain elements, biogeography focuses more specifically on how changes deeper in the Earth’s geological history have given rise to the diversity of life we see today. For instance, key themes in biogeography include an understanding of how plate tectonics and/or historical shifts in the global climate have influenced biological diversity through: a) the shaping of species’ dispersal patterns; b) immigration and emigration of individuals into and out of ecosystems; and c) the physiological and

reproductive fitness of different individuals or species in response to climatic change. Such factors in turn exert major controls on species demographics and effective population sizes, and the commonness or rarity of species (Sahney et al. 2010). At the same time, anthropogenic influences are creating scenarios (or “natural experiments”) where biogeographers can observe in real-time how species arise, the conditions of global change under which different species flourish or decay, and the key determinants of speciation, dispersal, and extinction (Kueffer 2015).

While the principles of biogeography tend to relate to “bigger picture” changes in the Earth systems, its conceptual tools and methods are widely applied to address real world conservation problems. Indeed, the principles of biogeography inform the conservation of species globally and are now well reflected in the emergence of the sub-discipline of “conservation biogeography” (Whittaker et al. 2005). For example, the “Island Theory of Biogeography” is arguably one of the most important guiding principles in conservation biology to date, as can be seen in the debate as to whether single-large or several-small conservation areas are the most effective in conserving biodiversity (commonly referred to as the “LOSS debate”; e.g., Tjørve 2010). Additionally, new tools are improving our capacity to link conservation biology with biogeography, including: a) global databases on species occurrences (e.g., the Global Biodiversity Information Facility (GBIF); Box 1), climate data (e.g., WorldClim), and geography and geomorphology (e.g., the Shuttle Radar Topography Mission (SRTM)) and b) free computing software that has high computational power (e.g., R Statistical Software, Maxent species modeling software). Coupled with rapidly expanding Internet access, these tools continue to increase our capacity to analyze and interpret current biogeographic trends, and use this information to inform conservation policy and practice. Furthermore, the open-access nature of such resources presents key opportunities for students (and citizens) to perform in-depth analyses that may be useful for conservation initiatives (Moritz et al. 2011).

### **OBJECTIVES**

In this exercise, you will: a) explore the spatial dis-



### **Box 1. Global Biodiversity Information Facility (GBIF)**

The Global Biodiversity Information Facility (<http://www.gbif.org>) is a global species occurrence database that relies on a large number of museums, universities, and research centers that publish information either collected or gathered by them, on the worldwide occurrence of species. A species occurrence is a record of: a) an individual of a species that has been observed; b) the method used to observe the particular individual and species; and c) the time and location of the observation. There are definitely sources of uncertainty surrounding species occurrence data. For instance, before satellite-based GPS coordinates were available, most of the records relied simply on qualitative descriptions of the location where samples were observed. Similarly, taxonomic identifications also present challenges. Some species can be identified and recorded under multiple scientific names (i.e., “synonym names”), and species identification is always challenging (e.g., some species are morphologically identical, and in other cases two individuals of the same species can be extremely different).

Accounting for such errors is critical in preventing systematic over- or underestimates for the total number of species. In addition to these challenges, another important limitation of species occurrence databases, such as GBIF, is that they contain “presence-only” information. This has two main implications. On one hand, GBIF and similar databases do not contain information about the abundance of the species, so the observation of one or two individuals of a species (i.e., “singletons” or “doubletons”, respectively) has the same weight as the presence of hundreds or thousands of individuals. On the other hand, occurrence data present an intriguing challenge, in that we often want to believe that such data also implicitly tell us something about the absence of species from certain locations.

Unfortunately, absences must be interpreted carefully since the absence of a species record in one location does not necessarily mean the species is not there. Researchers must consider why an area on the map has no records, and this can be either because: a) no biodiversity survey was ever conducted there; b) data from that location have not been uploaded to a particular database; c) the species may have been present in the location during a biodiversity survey, but was simply not observed; or d) the species is actually absent from that particular location and would not be observed even under the most intensive biodiversity sampling efforts. Other strengths and limitations of biodiversity databases have been discussed in the literature and include incomplete information, inaccurate locations, incomplete and/or biased sampling, among others (e.g., Otegui et al. 2013, Beck et al. 2013, 2014).

tribution patterns of one selected taxonomic group; b) examine the biogeographical changes this taxonomic group has experienced through geologic time; c) identify potential threats related to the conservation or spread of this taxonomic group; and d) identify and evaluate opportunities for conserving or managing this taxonomic group. In doing this, you will first review the core concepts of biogeography and then apply them to understand your group’s distribution using a global species distribution database. Additionally, you will explore the primary literature to expand your understanding of the processes that have led to the current species distribution patterns, and think critically about the possible conservation biology implications for your specific taxonomic group. Fundamentally, these goals will be met through the hands-on creation of species distribution maps, which will provide the most up to date information on your selected taxonomic groups along with the creation of an oral presentation and written report.

### **OVERVIEW**

For this assignment, the class will be divided into

teams of 3–4 students. During the next four weeks, each team will gather information via the GBIF species distribution database and create distribution maps for a given taxonomic group—hereafter referred to as your “taxon” (singular), or “taxa” (plural). Your taxon incorporates all the species within a particular taxonomic family or order. If information is available, you will also explore distribution maps of fossils for your taxon. All of these maps, along with information collected from primary literature, will help your team to form a number of testable hypotheses. Specifically, your team will derive hypotheses on: a) the geographical origin of your assigned taxon; b) the current distribution of the assigned taxon; c) the leading conservation risks for your assigned taxon; and d) the conservation mechanisms (including management) that can possibly address the current threats in the Anthropocene. At the end of the project, each team will share their results with the class in an oral presentation, and each team will submit a written scientific report based on their project (see rubric and Appendix I for more information on the report instructions). Over the course of the next four weeks, each team will be expected to work on their report outside of class time.



By Week 3, your explorations will allow you to answer the following questions from Box 2 (see below). As you proceed through the week activities, keep in mind that your answers to these questions will be included in the discussion section of your final report. Additionally, be aware that there are several leading journals in the fields of biogeography and biodiversity, including *PLoS ONE*, *Frontiers of Biogeography*, *Diversity & Distributions*, *Global Change Biology*, *Conservation Biology*, *Biological Conservation*, and *Journal of Biogeography*. These resources will also be helpful in developing the discussion for your report on the biogeography and conservation of your taxonomic group (also see the references).

## ACTIVITIES WEEK-BY-WEEK

### 1. Week One: Biogeography Data Acquisition From GBIF

During the first week of the assignment, your team will need to accomplish the following: a) choose your taxon from the list provided; b) download your dataset from GBIF; c) import your data into a working database (e.g., Microsoft Excel); and d) use the working database to derive a map of your taxon's global distribution. In doing so, you will also e) begin to document and explore the limitations of the data available on GBIF for your taxon.

#### 1.1 Select a Taxon

First, your team should select a taxon for your project and report your decision to your instructor. Based on

### Box 2. Questions for discussion

Section A: Understanding current distribution (*answer all* the following):

1. Describe the present distribution of your taxon in terms of biomes and habitat requirements. Also list the continents/regions where your taxon is present or notably absent; are there areas where the contrast-taxon is found but your focal taxon is absent? Or areas where both are absent?
2. Using your GBIF occurrence data, and your peer-reviewed literature review, speculate on which geographical area this taxon might originate from.
3. How might this taxon have arrived at its present distribution? Specifically, discuss what are the barriers and dispersal pathways that might have existed, which explain your taxon's current distribution. If fossil information is available in your maps, use it to discuss your observations. (Remember that barriers and continents may change in the time span of your taxonomic group. Compare the geographical history of your taxon with geological history. Take into consideration invasive species that can blur other patterns.)
4. Are there problems with the taxonomy for this taxon that limit or confound the understanding of its distribution? Explain your answer (e.g., consider the approaches in identifying species from your focal taxon and revisions made to its taxonomic classification).

Section B: Conservation biology of the taxon (*answer only two* of the following, based on which are most relevant for your taxon group):

1. Are there any endangered species in your taxon? Explain why or why not, based on your GBIF data and/or your literature search.
2. Is endemism a major factor that would affect the conservation status or potential management options for your taxon? Why or why not?
3. Would you expect the present-day distribution of your taxon to shift under climate change? How?
4. Does your taxon include any number of invasive species? Why or why not, and what are the likely mechanisms of invasion?

Section C: Conservation into the future (*answer only one* of the following):

1. Identify one region that you think should be protected, as a means to ensure the conservation of your taxon. In that area, identify two processes that threaten the survival of your taxon, and suggest mechanisms to deal with these threats (here, you may incorporate any real ongoing projects that may be in this area). If the threats that you are analyzing only apply to a subset of your taxon (e.g., one or a few species of a genus or family), then explain why you chose this particular species and region.
2. Identify one region where your taxon is considered a threat to an ecosystem. In this area, identify why a particular species from your taxon is a threat to the other species in the area, and suggest at least two mechanisms to manage and mitigate these threats (here, you may incorporate any real, ongoing projects that may be in this area). If only a subset of your taxon is a threat for the ecosystem/region, then explain why you chose this particular species and region.

Section D: Data limitations

1. Explain any major limitations associated with the methods that you used in this assignment.



a number of factors, including sufficient data coverage in the GBIF database, conservation importance, and public knowledge of these species, the following taxa have been identified as recommended options for this exercise:

1. Spine crawler mayflies (Ephemeroptera)
2. Pond turtles (Emydidae)
3. Narrow-mouthed frogs (Microhylidae)
4. Boa snakes (Boidae)
5. Viper snakes (Viperidae)
6. Megabats (Pteropodidae)
7. Horses (Equidae)
8. Beavers (Castoridae)
9. Opossums (Didelphimorphia)

After you have made your selection, you will begin by gathering general information about your taxon. In particular, you will need to focus on the types of ecosystems and habitats your taxon is associated with. How far does your taxon travel and disperse? Are there habitat types or environments that the taxon cannot cross, posing as dispersal barriers? Altogether, this information will help you to understand how your taxon might have become distributed throughout different parts of the world.

### 1.2 Download Your Data

Once your taxon has been selected and confirmed by your instructor, you will then download and process data on that taxon from the GBIF website (<http://www.gbif.org/>). Copy the suggested citation for your downloaded data from the GBIF website. Appendix II outlines these technical aspects and the step-by-step data acquisition process.

You will find that your GBIF dataset includes occurrence data for your taxon. Each row in the downloaded spreadsheet corresponds to an occurrence record with information about the observed species' taxonomy (order, family, genus, species), where the occurrence was observed (i.e., latitude, longitude, altitude, country, state), when the occurrence was observed (i.e., date of the record), and how that occurrence was recorded (e.g., live animal field observation, information from a museum collection, fossil collection). Re-read Box 1 for a

review of information about GBIF data.

Note here that GBIF occurrence data may be representative of the current distribution for a species, and/or a historic distribution (if fossil observations are available). Because the assignment is at a global scale, the distribution map you create from GBIF information will be reliable only if there have been enough chances (or adequate sampling effort) to allow for reliable surveys and records of the occurrence of your taxon around the globe. "Good sampling" implies many samples distributed widely across the expected distribution, without a strong sampling or reporting bias (such as intensive sampling only in one small part of the world).

### 1.3 Make a Map of Your Taxon

Your downloaded GBIF dataset is a snapshot of the information on the distribution of your taxon, made available from several museums, universities, and research centers around the globe. A key task is now to create maps that can help you to understand, and more importantly, visualize the distribution of your taxon (and the species in it) around the globe. These distribution maps that you are building do not intend to describe the probability of finding a given species, although some modern algorithms (e.g., Maxent, ENFA) can produce those types of maps using data from GBIF or similar sources. In this exercise, your group will be building maps similar to the maps used in the published biogeography literature, which are based on known distributions of the species (while also acknowledging information gaps in the process).

Computer templates that will help you build these maps are provided (see Appendix III), but first your team will need to organize your downloaded data in order to keep only the relevant pieces of information. Furthermore, in addition to a map that simply shows the occurrence data for your taxon, you will need to create supporting information. This will include: a) a secondary map showing the number of species of your taxon occurring in different parts of the world; b) a table that includes the names of the species in your taxon, along with information on the extent of their distribution, measured as the number of squares or grid cells each species occupies (each of which represents



a 10° latitude by 10° longitude area), and lastly c) a table that relates valid species names in your taxon with their synonyms. Appendix III explains in detail how to use a simple database management program (namely, Microsoft Excel) to summarize the thousands of records you downloaded from GBIF, into a smaller dataset that contains only the required information to build the maps and tables.

#### **1.4 Using a Contrast Map to Document Sampling Bias**

This is a good stage in the assignment to begin documenting some of the limitations of biogeographic data. Specifically, at this stage you may start to notice evidence of sampling bias, something very common in biodiversity and biogeography data (Zhang et al. 2014). Presence-only data make it hard to distinguish sampling bias from real absences (see Box 1). With respect to sampling bias then, your group must think critically as to whether or not the occurrences in your downloaded GBIF database are truly representative of the global distribution of the species.

For example, it is often the case that areas further from roads are harder to sample, developed countries have more resources for reliable sampling, and countries more interested in connecting to the GBIF are better represented. Also, sampling bias certainly occurs across taxa, with charismatic species generally having larger and more consistent sampling efforts. Consider the panda bear, *Ailuropoda melanoleuca* (Ursidae), as an example: a great deal of resources are available for protecting this species, and there are intensive monitoring programs in place. In contrast, the red panda, *Ailurus fulgens* (Ailuridae), is a less charismatic species and as a result receives less funding and support for monitoring. In turn, the red panda is considerably less well evaluated in biodiversity databases as compared to the panda. Additionally, zoo records can commonly be included in databases which can bias your datasets, and even subtle spelling mistakes in species names or typos in coordinates may lead to data biases. For example, if you find an elephant living in Toronto or a polar bear in the Caribbean, they are likely taxonomic or coordinates typos or specimens in zoos.

One way to distinguish areas with real absences, vs. areas

with no sampling effort, is to look for species that can be sampled using methods similar to the ones used for the taxon of interest. For instance, if you are studying a particular group of bats (such as vampire bats), how can you use information collected for other bat species? One way is by assuming that bat sampling methods (such as mist nets) are likely to capture most bat species, but not other species with different characteristics such as mice or cats. Therefore, a contrast map, a map showing the distribution of all the bats recorded in the database will provide a rough estimate of the distribution of “samples” (places where bats have been collected). For example, an area that has both a high density of bats (i.e., an area with several samples) but is lacking in vampire bats would indicate a real absence of vampire bats: in other words, no species occurrence records despite a large sampling effort are much more likely to be real absences. The contrast map will provide a contrast to compare with the distribution of your taxon of interest.

Since this assignment is at a fairly coarse spatial and taxonomic scale, the presence-absence sampling pattern should be good enough to give an idea of potential sampling bias. In any case, the conclusions gathered from the map will be revisited during the second and third week. Discuss with your team and your instructor which taxon may be a good option to use as a contrast group to compare with your species of interest.

#### **1.5 Week One Outcomes**

At the end of week one, your group should have:

- a brief note about the primary habitat requirements and common dispersal characteristics of species representative of your taxon;
- a table of occurrence for your focal taxon, obtained from the GBIF database (file occurrences.csv);
- a table that relates valid species names in your taxon with their synonyms (Appendix III, Step 3);
- maps of your taxon’s distribution as obtained using the “map” spreadsheet in the Excel file “dataFrame.xlsx”;
- two secondary maps showing the locations of a) the number of extinct (fossil) species and b) extant species of your taxon, both created using the “map” spreadsheet in file “dataFrame.xlsx” (for the fossil map, choose the tag “richness” and fossil



“TRUE” in the top selectors in the spreadsheet, for the extant species use the same tag but fossil “FALSE”);

- a table with the scientific names of the species in your taxon, along with information on the number of squares where each species is represented (created using spreadsheet “taxaAbundance” in spreadsheet “dataFrame.xlsx”; where each square represents a 10° latitude by 10° longitude area);
- contrast map obtained directly from GBIF showing the distribution of the extant taxon that will be used to assess the sampling bias (Appendix II, Part 2), and your team’s assessment of the level of bias likely to apply to your taxonomic group.

## 2. Week Two: Data Exploration

For the second week, your group should come prepared with the outcomes from the first week and be ready to: a) explore your data; b) evaluate its consistency; and c) gather additional information you may need to answer the questions of this assignment (outlined above in Box 2). In general, by addressing these questions, you will begin to link ideas about how multiple different ecological, evolutionary, and landscape-level processes may have shaped the distribution of your taxon. Review the third weeks’ activities to have a better idea of the type of information you will need and then distribute the tasks among your group members accordingly.

### 2.1 Exploring Your Data

Now that you have your basic study tools from Week 1, you can start creating distribution maps for the different species in your group. In this step, use your curiosity to create some questions and answer them. The first questions can (and perhaps should) be focused on exploring and learning Excel functionality in order to map the taxonomic distributions you compiled in Week 1. Specifically, how can you use the filters to map fossils and living specimens?; How can taxonomic richness be mapped?; Where are the actual continents in your coarse resolution map?; Or where are the archipelagos? Once these more technical questions are addressed, you can start exploring whatever scientific ideas catch your interest: for example, which species have larger latitudinal range? Which species occupy a greater

area in the world? Are these species distributed across the entire world? Are the species with smaller ranges constrained to a particular geological/geographical part of the world? Also, you can now examine the richness maps that indicate the number of species (either fossil or extant) in different parts of the world and ask yourself whether they show similar patterns.

### 2.2 Evaluating Data Consistency

Start by identifying potential gaps in your occurrence table downloaded from GBIF. For instance, if there is an area in South America with no occurrences of your taxa, and no occurrences of your contrast group either, search the scientific literature for papers focused on your taxon in those countries, and use these sources to confirm that your taxon is indeed not likely present in these areas. Also, by this stage your group should have evaluated whether or not your taxon has synonyms; if synonyms were identified, review the scientific literature and evaluate the distribution of those synonymous species. If you cannot find any scientific literature on your taxon within a region that your GBIF data also suggest does not contain your taxon, you have indeed identified a likely real absence in the distribution of the taxon. If you find a paper about your taxon (or synonym) in a particular region, you will have to add this record in your Excel database. Before adding the information, make sure that the species name you are adding is indeed valid (remember to consult your list of valid taxon names and their synonyms obtained by the GBIF database during the first week). These gaps should be included in your final report, along with the additional literature you have found.

### 2.3 Gathering Relevant Information

Different taxa have widely different evolutionary histories and face a variety of conservation problems. Use the discussion questions in Box 2 to direct your readings. The first piece of relevant information that you will want to document is the evolutionary history of your taxon. Specifically, attention should be given to the ‘age’ of your taxon, the current understanding of its evolutionary origins, the possible pathways it followed to disperse into its current distribution, potential extinction events that occurred in the past,



or other events that led to the isolation of a once widespread species (i.e., a need for your taxon to retreat to refugia). In sum, this information collectively should give your team an idea about how your taxon's current-day distribution patterns came to be. Use your map in `dataFrame.xlsx` to generate fossil maps and explore the information obtained from GBIF.

In addition to your taxon's evolutionary history, your team should also document patterns of your taxon's endemism: any major restrictions on the distribution that confine your taxon to a particular geographical area. This is highly relevant for some taxa (especially those that do not disperse particularly well) and can be very helpful in understanding the origin of a taxon. Your team should also note whether or not your taxon contains any invasive species; identifying invasive species, particularly those facilitated by human movement, is important since such species can introduce complexity when interpreting spatial patterns and their mechanisms. Your team should attempt to discern instances of endemism or invasiveness, by using the table that counts the number of squares (each of which represents a 10° latitude by 10° longitude area), the taxon of interest occupies in the Excel map (see "taxaAbundance" spreadsheet in the "dataFrame.xlsx" file); endemic species should occupy a disproportionately small number of cells, while invasive species should occupy a disproportionately large number of cells. You may wish to reproduce these Excel maps for specific species to show the scale at which they are endemic or invasive to a region. Organize your findings of this and the previous week in a text that can be used later for the introduction of your final report.

### 2.4 Week Two Outcomes

At the end of week two, your team should have:

- a compilation of literature addressing the age (or hypothesized age) of your taxon, and of any particular subgroup you may find important to explain how the focal taxon arrived at its present-day distribution;
- all the data products from week one (see Section 1.5), supplemented with the data acquired during this week, including notes on the data consistency and invasive/endemic species; and
- an expanded written description of the taxon to

include in the final report—this should provide you and the reader with an understanding of the potential habitat requirements and barriers, including the dispersal mechanisms involved.

## 3. Week Three: Data Analysis and Questions

At the beginning of the third week, your team should come prepared with complete background descriptions of your taxon, including data tables, distribution maps, and written background information. You should begin to organize these data in such a way that they are useful for answering the discussion questions (see Box 2). Each member of your team will come to this week's session with your initial ideas surrounding potential answers and any questions you may have for your instructor to the discussion questions and questions for your instructor.

During the third week, your team will begin to prepare your final report, will answer the specific questions (see Box 2), which should be addressed in the report's discussion section, and will prepare your 12-minute presentation (note: time requirements may be altered based on your instructor's discretion).

### 3.1 Discussion Questions

The questions in Box 2 are designed to summarize the biodiversity and biogeography information that you have compiled on your focal taxon and are a guideline for the discussion section of your report (GBIF occurrence data, distribution maps and other tables, literature review). The questions are divided into three sections; the first section (A) focuses on biogeographic patterns, in particular the distributional map, as well as guidelines to understanding your map. The second section (B) introduces questions that may or may not be relevant to your particular taxon. Your first task in section (B) is to *identify and answer two* of these questions relevant to your taxon. In the third section (C), you will *identify and answer only one* of the questions to integrate the information collected previously into the context of conservation threats faced by your taxon. Finally, in the fourth section (D) you will discuss the potential limitations of your study. See the rubric for some extra details on each question. Drawing upon the Excel results and skills you developed in Week 1, create maps and



tables than can help you to answer those questions, and present them in the results section, with a brief description of the most important patterns you observe that are relevant to answering the discussion questions (refer to the rubric and Appendix I). In the discussion, provide detailed and fully developed answers and reference scientific literature as appropriate.

### 3.2 Preparing Your Presentation

Each team member will be required to speak, and provide information that gives either background, answers specific questions or summarizes the overall report on this taxon, as appropriate. As a team, you are free to choose your presentation format, but it is expected that your presentation will cover the assignment topics (in particular the discussion questions) and integrate your understanding of your focal taxon's distribution and ecological requirements. If you do not think you will have enough time to explain everything, choose the most relevant points of each question that you want to share or discuss; for instance, you may choose to explain distribution patterns for a smaller number of the most representative species within your taxon that are poorly represented, endemic, or highly threatened species. Whatever you choose, make sure to explain why they were chosen briefly your decision. Detailed guidelines on the presentation are in the rubric and Appendix I.

### 3.3 Looking Ahead

During the fourth week each team will: a) present their results to the whole class in a 10-minute presentation; b) receive feedback and suggestions to improve their report; c) incorporate that feedback into their report; and d) submit one single team final report for their assigned taxon. It is also advisable to contextualize and discuss the findings of the other teams compared to the findings of your team in your final written report. Specific guidelines for the report (length of each section, requirements, etc.) can be found in the rubric and Appendix I.

### 3.4 Week Three Outcomes

At the end of the third week, your team should have:

- a working draft of your report that demonstrates:

- a clear understanding of the distribution of your taxon
  - a clear understanding of the processes that shaped your taxon distribution
  - a clear understanding of potential implications of your taxon's distribution for conservation
  - a discussion section that addresses the questions selected from Box 2.
- A 10-minute presentation which will be presented in Week 4

## 4. Week Four: Project Presentations

At the beginning of the fourth week, your team should come prepared with your presentation, as well as a draft of your final written report. The final class period will be devoted completely to presentations, where each team's analysis and recommendations will be presented to the class.

### 4.1 Week Four Outcomes

At the end of week four, your group should have:

- presented your analysis to the class
- a clear understanding of the strengths and weaknesses of your analysis
- a plan to incorporate student and instructor feedback received during your presentation, into your final report prior to its final submission.

## Final Report

You will have some time after your presentation to make any last-minute revisions or changes in their answers/report based on information raised during the final presentation. Review the schedule provided by your instructor.

## REFERENCES

- Beck, J., L. Ballesteros-Mejia, P. Nagel, and I.J. Kitching. 2013. Online solutions and the 'Wallacean shortfall': what does GBIF contribute to our knowledge of species' ranges? *Diversity and Distributions* 19(8):1043–1050.
- Beck, J., M. Böller, A. Erhardt, and W. Schwanghart. 2014. Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. *Ecological Informatics* 19(2014):10–15.



- Crutzen, P.J. 2002. Geology of mankind. *Nature* 415(6867):23–23.
- Douglas, J. 1978. Biologists urge US endowment for conservation. *Nature* 275(5676):82–83.
- Kueffer, C. 2015. Ecological novelty: towards an interdisciplinary understanding of ecological change in the Anthropocene. Pages 19–37 in H. Greschke and J. Tischler, editors. *Grounding Global Climate Change*. Springer, Dordrecht, Netherlands.
- Ladle, R. and R.J. Whittaker. 2011. *Conservation Biogeography*. Wiley, Hoboken, NJ.
- McKinney, M.L. and J.L. Lockwood. 1999. Biotic homogenization: a few winners replacing many losers in the next mass extinction. *Trends in Ecology & Evolution* 14(11):450–453.
- Millennium Ecosystem Assessment. 2005. *Ecosystems and Human Well-being: Synthesis*. Island Press, Washington, DC.
- Moritz, T., S. Krishnan, D. Roberts, P. Ingwersen, D. Agosti, L. Penev, M. Cockerill, and V. Chavan. 2011. Towards mainstreaming of biodiversity data publishing: recommendations of the GBIF Data Publishing Framework Task Group. *BMC Bioinformatics* 12(Suppl 15):S1.
- Otegui, J., A.H. Ariño, M.A. Encinas, and F. Pando. 2013. Assessing the primary data hosted by the Spanish node of the Global Biodiversity Information Facility (GBIF). *PLoS One* 8(1):e55144.
- Pachauri, R.K., et al. 2014. *Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the fifth assessment report of the Intergovernmental Panel on Climate Change*. IPCC, Geneva, Switzerland.
- Richardson, D.M. and R.J. Whittaker. 2010. Conservation biogeography—foundations, concepts and challenges. *Diversity and Distributions* 16(3):313–320.
- Sahney, S., M.J. Benton, and P.A. Ferry. 2010. Links between global taxonomic diversity, ecological diversity and the expansion of vertebrates on land. *Biology Letters* 6(2010):544–547.
- Soulé, M.E. 1985. What is conservation biology? *BioScience* 35(11):727–734.
- Soulé, M.E. 1986. *Conservation Biology: The Science of Scarcity and Diversity*. Sinauer, Sunderland, MA.
- Tjørve, E. 2010. How to resolve the SLOSS debate: lessons from species-diversity models. *Journal of Theoretical Biology* 264(2):604–612.
- Whittaker, R.J., M.B. Araújo, P. Jepson, R.J. Ladle, J.E.M. Watson, and K.J. Willis. 2005. Conservation biogeography: assessment and prospect. *Diversity and Distributions* 11(1):3–23.

**RUBRIC****Team members (initials):**

<b>TITLE PAGE (1 PAGE)</b>					
Indicate primary responsibility of each student (sections and questions). Each group member must sign the report beside their name accepting the division of labor as indicated.	Mandatory				
<b>INTRODUCTION (400 WORDS; ~1.5 PAGES)</b>					
<b>Criteria</b>	<b>Absent or irrelevant</b>	<b>Present and relevant</b>	<b>Relevant, concise, and well documented</b>	<b>Mark</b>	<b>Comments</b>
The topic of the study and primary research questions are well presented and easily flows for an academic reader.	0	1.5	3		
The main characteristics of the taxon (habitat, dispersion, age, etc.) are neatly and succinctly explained.	0	1	2		
Conservation issues of the species are discussed, if needed, with emphasis in particular areas.	0	1	2		
An overview of the data analysis is explained.	0	0.5	1		
<b>MATERIALS &amp; METHODS (500 WORDS; ~2 PAGES)</b>					
<b>Criteria</b>	<b>Absent or irrelevant</b>	<b>Present and relevant</b>	<b>Relevant, concise, and well documented</b>	<b>Mark</b>	<b>Comments</b>
The approach to obtain the data and the sources is clearly presented.	0	1.5	3		
General explanation of how every single map and table in the document (including appendices if present) was constructed.	0	1.5	3		



<b>RESULTS (INCLUDING TABLES, FIGURES, MAPS; ~3-4 PAGES OF TEXT; ~1000 WORDS)</b>					
<b>Criteria</b>	<b>Absent, irrelevant, or inaccurate</b>	<b>Partly relevant or partly accurate</b>	<b>Fully relevant and fully accurate</b>	<b>Mark</b>	<b>Comments</b>
Figures/tables: It is easy to understand the message from the figures and tables (figure is appealing, units are present, properly labeled and captioned).	0	3	5		
Figure/table: The figures and tables help to support the message in the document.	0	3	5		
The most important patterns presented in the figures and tables, and relevant to the discussion, are properly described.	0	2	4		
The results are easy to read and no more tables or figures seem required to supplement the message of the document.	0	2	4		
<b>DISCUSSION: (~1800 WORDS; ~7 PAGES MAX)</b>					
<b>Criteria (by lettered parts)</b>	<b>Absent, irrelevant, or inaccurate</b>	<b>Partly relevant or partly accurate</b>	<b>Fully relevant and fully accurate</b>	<b>Mark</b>	<b>Comments</b>
A. Description of the current distribution of the taxon.	0	4	6		
A. Discussion of the area of origin of the taxon.	0	3.5	5		
A. Discussion of the potential dispersal patterns and barriers of the taxon and how they lead to the current distribution.	0	4.5	8		
A. Discussion of the challenges in the taxonomy of the group.	0	2	3		
B. <i>Two questions selected from B list help to understand the challenges faced by the taxon.</i>	0 (Irrelevant questions)	1 (Partially relevant questions)	2 (Two relevant questions)		Not needed to explicitly address this point.



B. The reasons behind the presence or absence of endangered species are discussed.	0	3	4		
B. Endemicity patterns are used to logically analyze conservation priorities.	0	3	4		
B. The potential impacts of climate change on the taxon are analyzed.	0	3	4		
B. Different mechanisms of invasion are analyzed in the particular context of these species.	0	3	4		
C. An area of the world is identified, described and the reason for its selection explained.	0	1	2		
C. Two threats and two solutions for them are justified, described and discussed.	0	3	4		
D. Limitations: The limits of the study are clearly presented and the implications for the results briefly described.	0	2	3		
<b>CONCLUSION (~250 WORDS; ~1 PAGE)</b>					
<b>Criteria</b>	<b>Absent, irrelevant, or inaccurate</b>	<b>Partly relevant or partly accurate</b>	<b>Fully relevant and fully accurate</b>	<b>Mark</b>	<b>Comments</b>
Complete and succinct characterization of the distribution of the group, potential origin, barriers and pathways, conservation challenges and relevance for conservation.	0	4	6		



<b>REFERENCES CITED (~250 WORDS; ~1 PAGE)</b>					
<b>Criteria</b>	<b>Absent, irrelevant, or inaccurate</b>	<b>Partly relevant or partly accurate</b>	<b>Fully relevant and fully accurate</b>	<b>Mark</b>	<b>Comments</b>
Citations used throughout the document are in a consistent and recognized format.	0	1	2		
All sources cited in report are listed in References section using a consistent and recognized format.	0	1	2		
<b>FORMAT</b>					
<b>Criteria</b>	<b>Absent, irrelevant, or inaccurate</b>	<b>Partly relevant or partly accurate</b>	<b>Fully relevant and fully accurate</b>	<b>Mark</b>	<b>Comments</b>
Typed, 1.5 or double-spaced & appropriate length. Clear & concise presentation of text (i.e., text flows logically and is coherent, correct spelling, proper grammar and structure overall).	0	1	2		
<b>DATABASE</b>					
Spreadsheet file used to analyze the data.	Mandatory for report to be marked and for results to be verified				
<b>GROUP REPORT (TOTAL)</b>	<b>0</b>		<b>85</b>		



ORAL PRESENTATION (INDIVIDUAL - 15 POINTS)	MAX	STUDENTS (INITIALS)			
Presentation is easy to follow with clear organization (i.e., specific intro and conclusion), no ideas “out of place”, and no irrelevant information.	3				
Content is directly relevant to the assigned topic, and all supporting evidence is accurate, high quality, and directly used to evaluate specific research questions.	2				
Students display a complete understanding of the subject matter, clearly explain the research questions, supporting evidence, and end with a clear and comprehensive take home message.	3				
Delivery makes the presentation compelling, speakers are confident, language is clear and easy to follow.	2				
Presentation includes clear, relevant, and aesthetically pleasing aids which are directly related to topic; all slides have appropriate text and precise, relevant info.	3				
Timing is appropriate and allocated sufficiently across all different parts of the presentation.	2				
	Total				



## APPENDIX I. WRITTEN REPORT

You will write your report following standard scientific paper guidelines. Below are some general suggestions, but review the rubric for specific requirements. For more information on scientific writing, see the NCEP module *Scientific Writing*, available for download at [ncep.amnh.org](http://ncep.amnh.org).

### Introduction Guide

The *Introduction section* must give the reader basic knowledge to be able to understand the whole purpose of the paper. You must make the introductory statement clear, giving the context for the study and explanation of the perspective. What are the key questions you ended up exploring with these data? Why did you take this approach? What information is already out there (what literature) to help you understand the knowledge that you are building?

The introduction should also describe the most important characteristics of the taxon you are working with, where it is and has been found, what types of habitats and the climatic conditions necessary for its survival. Also, some mention of its dispersal capabilities would be important to mention to allow the reader to understand your rationale and discussion below. If appropriate, you may also wish to briefly build a case for conservation of your focal taxon by presenting some of the threats that may limit its survival.

### Methods Guide

The *Methods section* is a succinct and brief description of the most important steps that you took to get your results, so that anybody can understand how you obtained your results and, if they wanted to, could replicate the study. Here, you have to explain how you got the results that you are presenting, without specific details of the steps you used to get them. For instance, you have to refer to the fact that you used GBIF data information, and you have to cite it following their suggestion, but you don't have to explain how to download the data.

### Results Guide

In the *Results section*, you should write about and highlight the major results that you identified from your

data analysis. In addition to summarizing your results within the text, the data should be presented in the form of maps, tables, charts, graphs, figures, histograms, or any other tool that can help to synthesize and visualize the most important findings that you are trying to communicate. You may want to include a map describing the number of species (within your focal taxon) in each area of the world, and also some maps for those species that show the most interesting distribution patterns (those patterns that you will most likely discuss in your next section). Don't forget to include a legend (e.g., color scale) in your maps and captions to describe your figures and tables.

### Discussion Guide (Questions To Address)

The *Discussion section* should include answers to the specific questions previously provided in Box 2, along with the implications of the results. These questions should be answered in scientific writing format rather than short answer format (e.g., don't list the questions/answers, but instead write a narrative that provides a detailed and fully developed discussion of your answers). Discuss possible explanations for biogeographic distributions and potential issues related to conservation. Finally, what gaps existed in the literature or in your data that present limitations to your analysis? In some cases, you may need to restate some of the ideas that you presented in the results, which is okay, but in this section you will need to add references to put your comments in context of previously published literature.

Optional: If you prefer and receive approval from your instructor, you may write a section combining results and discussion. In that case, you can link to one of your figures or tables that can help answer the particular question you are addressing. Be aware that some questions will need specific results, while others may require a more comprehensive analysis of your general results.



---

## Conclusions Guide

The *Conclusions section* will address the overall importance of your taxon, its origins, past and present distributions, ecological challenges (barriers/pathways), and significance for conservation.

## Reference Guide

All references used to support the introduction, methods, and discussion sections of the paper must be cited in proper order at the end. Only appendices (optional, if you decide to not embed figures and tables within the text) will appear after this section. Try to limit your literature to peer-reviewed journals. If you are unsure of the reliability of sources, consult with your instructor.

Each reference source should be listed alphabetically by author and provides sufficient information on it so that any reader will be able to retrieve it and verify the statements made in the paper. The format of the reference section is less important than the requirement to be consistent in formatting throughout. One suggested way of proper referencing is to use the standardized APA or similar format: [www.apastyle.org/learn/tutorials/basics-tutorial.aspx](http://www.apastyle.org/learn/tutorials/basics-tutorial.aspx).



## APPENDIX II. RETRIEVING DATA FROM GLOBAL BIODIVERSITY INFORMATION FACILITY (GBIF)

GBIF is organized in four sections: Occurrences, Species, Datasets, and Data Publishers. The four sections are linked, but you will work only with the first two ones.

1. Downloading occurrences:
  - a. Go to <http://www.gbif.org>.

GBIF | Global Biodiversity Information Facility

# Free and open access to biodiversity data

OCCURRENCES SPECIES DATASETS PUBLISHERS RESOURCES

Search

WHAT IS GBIF? ABOUT GBIF CANADA

<p>Occurrence records</p> <p><b>874,743,589</b></p>	<p>Datasets</p> <p><b>37,393</b></p>	<p>Publishing institutions</p> <p><b>1,135</b></p>	<p>Species</p> <p>Learn more about the number of species covered by data in GBIF.org.</p>
---	--------------------------------------	--	---

Call for nominations opens for 2018 GBIF Young Researchers Award  
30 November 2017

The future of ragweed: making more Europeans sneeze  
4 December 2017

Job opportunity for Data Analyst at the GBIF Secretariat  
23 November 2017

Job opportunity: Data Administrator at the GBIF Secretariat  
27 November 2017

- b. Select occurrences.
- c. The screen is now divided in three main parts, the menu on the top, the filter menu on the left, and the occurrences report on the right. While you apply your filters (left menu), you will notice that the number of records that match the result changes (“search occurrences” in top middle section) will change.



Scientific Name	Country	Coordinates	Basis Of Record	Month & Year
Cygnus cygnus (Linnaeus, 1758)	Norway	58.3N, 7.7E	human observation	2017 January
Cygnus olor (Gmelin, 1789)	Norway	59.1N, 10.4E	human observation	2017 January
Accipiter gentilis (Linnaeus, 1758)	Norway	59.9N, 10.8E	human observation	2017 January
Sturnus vulgaris Linnaeus, 1758	Norway	67.5N, 12.1E	human observation	2017 January
Parus major Linnaeus, 1758	Norway	59.0N, 10.1E	human observation	2017 January
Turdus pilaris Linnaeus, 1758	Norway	59.9N, 10.8E	human observation	2017 January
Parus major Linnaeus, 1758	Norway	59.9N, 10.7E	human observation	2017 January
Streptopelia decaocto (Frisvoldszky, 1838)	Norway	58.1N, 8.0E	human observation	2017 January
Anas crecca Linnaeus, 1758	Norway	59.9N, 10.7E	human observation	2017 January
Surnia ulula (Linnaeus, 1758)	Norway	59.2N, 9.7E	unknown	2017 January
Falco tinnunculus Linnaeus, 1758	Norway	59.2N, 11.0E	human observation	2017 January
Periparus ater (Linnaeus, 1758)	Norway	60.0N, 11.1E	human observation	2017 January
Aythya fuligula (Linnaeus, 1758)	Norway	58.5N, 6.3E	human observation	2017 January

- d. Add a *Scientific name* filter: Search for your taxon (order, subclass, superfamily, genus, species, etc.). Tip: use the dropdown list (you may need to do an internet search to figure out which groupings to select at each level). BE CAREFUL! A same genus, species, or family name can be used in different kingdoms.

Search all fields

Simple Advanced

Record License

Scientific Name

boidae

**Boidae Family**  
Animalia > Chordata > Reptilia > Squamata

Location

Year

Month

Dataset

Country

Issue

Media Type

Publisher

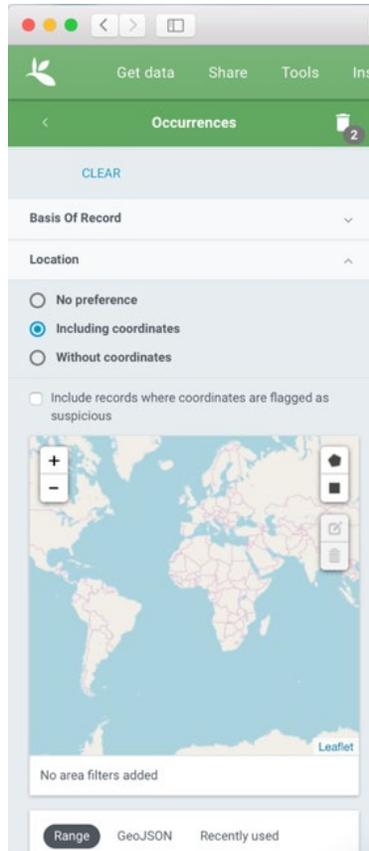
Institution Code

Collection Code

Catalogue Number



- e. Add a location filter, to filter out the occurrences without coordinates.



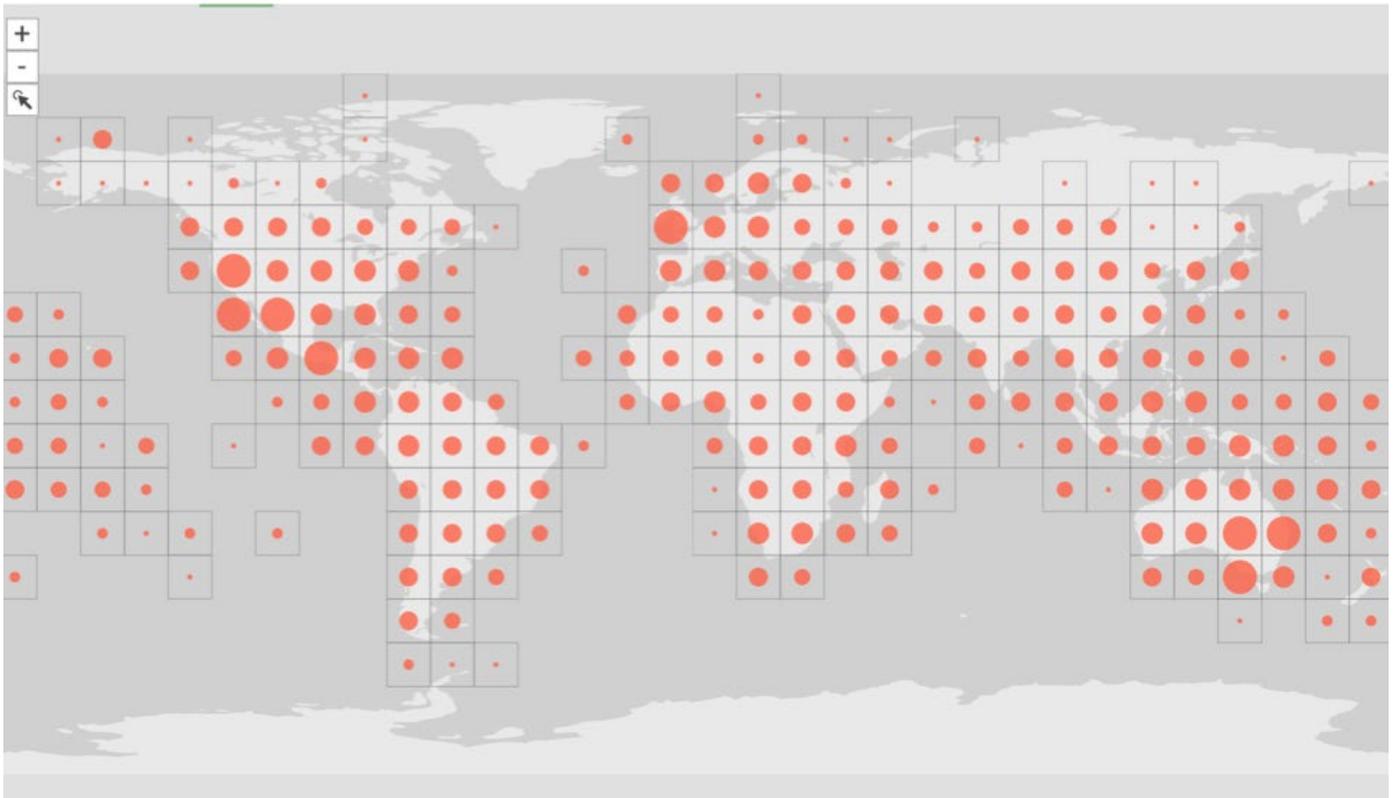
- f. Add other filters that you might consider useful.
- g. Press the download button in the top-center. To download the data, you will need to create an account. Follow the steps indicated in the webpage.
- h. Once your account is created and you press the download button, you will be able to choose the format. Simply press the CSV button. Take some time to review the information provided in the webpage.



- i. After you press the download button, the server (the computer at GBIF) will begin to gather information from the different servers around the world that can contain relevant information for your query. Once it is ready, you will receive an email saying that you can download your file. The email contains a link and a reference! Keep that email to add the reference later to your report.
  - j. Click on the provided download link in the email and automatically the file will download (often as a zip file). Unzip and save it with this name: *occurrence.txt* (the default name will be *occurrence.csv*, try to change to *occurrence.txt*. See suggestions below. Talk with your instructor if you have problems doing it). You may receive a prompt from your computer warning you that you are changing the extension of your file. Ignore the warning and continue.
2. It may take up to 15 minutes for the data to download from the servers, so take this time to review the general information of your taxon from GBIF by clicking on the other categories in the top section (e.g., map). These sections are especially useful to determine what your contrast group will be.
- a. Repeat steps (a) to (f) from the previous section, using another taxon (e.g., your contrast group, anything that interests you). Feel free to explore other orders, families, genera, species or any other taxonomic levels.
  - b. With your contrast taxon, browse through GBIF information as you did with your focal taxon. There is no need to download the GBIF data for the contrast taxon.

SEARCH OCCURRENCES | 3,258,009 WITH COORDINATES

TABLE GALLERY MAP SPECIES DATASETS DOWNLOAD



- c. At least for your contrast taxon, select the map view, take a screenshot and save it for future inclusion in the final report.



## APPENDIX III. IMPORT YOUR DATA INTO A DATABASE (EXCEL AND R)

### Excel Instructions (Adaptable to Other Similar Software)

The following instructions have been developed in Excel for Mac 2011, so certain details may differ. All of the functions are likely available in more recent versions but may be located in different menus or ribbons. Similarly, alternative programs such as LibreOffice, Numbers or Google Sheets will likely offer the same capabilities, but the location of the functions will differ compared to the steps detailed here.

*Warning:* Early versions of MS Excel (i.e., Excel 2003 and earlier) are unable to handle more than 65,536 rows of data. Newer versions can handle up to 1,048,576 rows. When you query GBIF, if you have more occurrences than the maximum number of rows permissible through the version of Excel being used, then another software may need to be used with the file.

Before using this document, be aware that you need:

- The *occurrence.txt* file downloaded from GBIF (see Appendix II). This file provides the species distribution information.
- The *dataFrame.xlsx* file provided by your instructor. This file contains a simplified base map of Earth and formatting information that will allow you to superimpose your data on to it at later stages of the assignment.

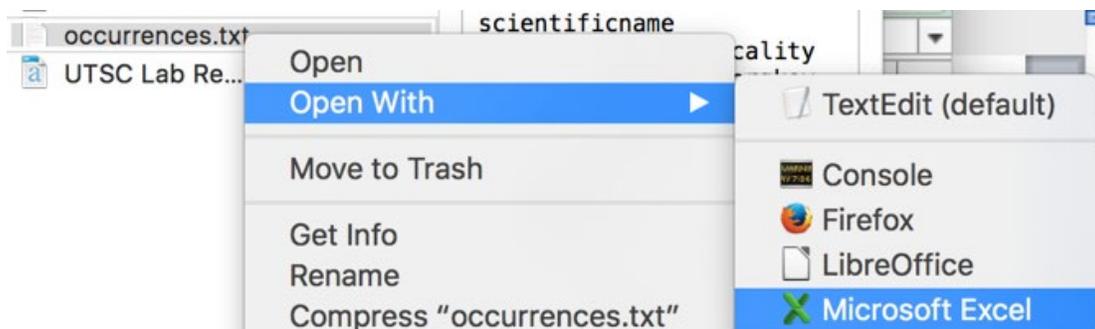
#### Tips: General suggestions when dealing with large datasets

- Keep a backup of your raw data.
- Never sort the raw data: if you don't do it properly, you will shuffle your data and it will no longer make sense.
- Save different versions of your file with each of your intermediate steps.
- Confirm the results you obtain in each step: it is very hard to track the source of data issues at the very end following extensive analysis, so confirming results at each step is advisable.

### Step 1. Open the *occurrence.txt* file downloaded from GBIF

This is likely the least interesting step, but can potentially be the most frustrating and discouraging because of the small computer formatting-related details. Specifications at this stage may also change from computer to computer owing to (for example) your browser configuration. At this stage, do not hesitate to ask for help if you cannot open the file. It is expected that you may need a bit of assistance during this step.

From Finder (on Mac OS x) or Explorer, you can try this option:





From Excel, use the following command: File > Open > Choose your file.

Since your GBIF data is downloaded as a text file, it is likely that you will see the “Text import wizard” window:

Text Import Wizard - Step 1 of 3

The Text Wizard has determined that your data is Delimited.

If this is correct, choose Next, or choose the Data Type that best describes your data.

Original data type

Choose the file type that best describes your data:

- Delimited - Characters such as commas or tabs separate each field.
- Fixed width - Fields are aligned in columns with spaces between each field.

Start import at row:  File origin:

Data preview

Preview of file Macintosh HD:Users:caam:Desktop:caam:TA:ES...:occurrences.txt.

	gbifid	datasetkey	occurrenceid	kingdom	phylum	class	order	family	genus	spe
1	1054805926	78122332-6315-41bd-914b-e9c1342d9093	urn:occurrence:Arctos:UWBM							
2	1054805944	78122332-6315-41bd-914b-e9c1342d9093	urn:occurrence:Arctos:UWBM							
3	1038342509	50c9509d-22c7-4a22-a47d-8c48425ef4a7	http://www.inaturalist.org							
4	1042789251	5e1d2d54-f5db-43ac-87c5-55a9f79ac718	9155	Animalia	Chordata	Rep				
5	1042789693	5e1d2d54-f5db-43ac-87c5-55a9f79ac718	9255	Animalia	Chordata	Rep				
6										

Cancel < Back Next > Finish

In this case, select “Delimited”, press “Next >”.

Now you have to choose the character that will be used to delimit (or separate) the cells from one another. Despite the file extension originally being “.csv” (which means “comma separated values”) when you downloaded it from GBIF, the values are actually delimited by tabs. So check that option only in the “import wizard”.





Finally, review and familiarize yourself with your file to be sure everything is displaying properly in different columns. It is usually a good idea to freeze the first row. (In Excel, follow this command: Layout > Freeze panes > Freeze top row.)

## Step 2. Cleaning the data

Keep only the useful columns that include information about what was collected (taxonomy), how, when, and where it was collected. You may not use all of these columns, but they may help you to understand your results later. In summary, your working data should retain only the following columns:

TYPE OF DATA	WHAT WAS OBSERVED?					
Column	A	B	C	D	E	F
Field	family	genus	species	infraspecific epithet	scientific name	
Content	Taxonomy (Family)	Taxonomy (Genus)	Currently valid scientific name	Taxonomy (infraspecies level)	Scientific name as originally reported	Leave this column empty

**Note:** If the data in the columns entitled “species” (Column C) and “scientific name” (Column E) do not match, it means that there may be a synonym that you must be aware of. You will identify these in the next step.

TYPE OF DATA	WHERE WAS IT OBSERVED?				WHEN WAS IT OBSERVED?	HOW WAS IT OBSERVED?
Column	G	H	I	J	K	L
Field	countrycode	locality	decimallatitude	decimallongitude	year	basisofrecord
Content	Country	First territorial level inside a country	Latitude	Longitude	Year when the sample was collected	Fossil, direct observation, unknown?

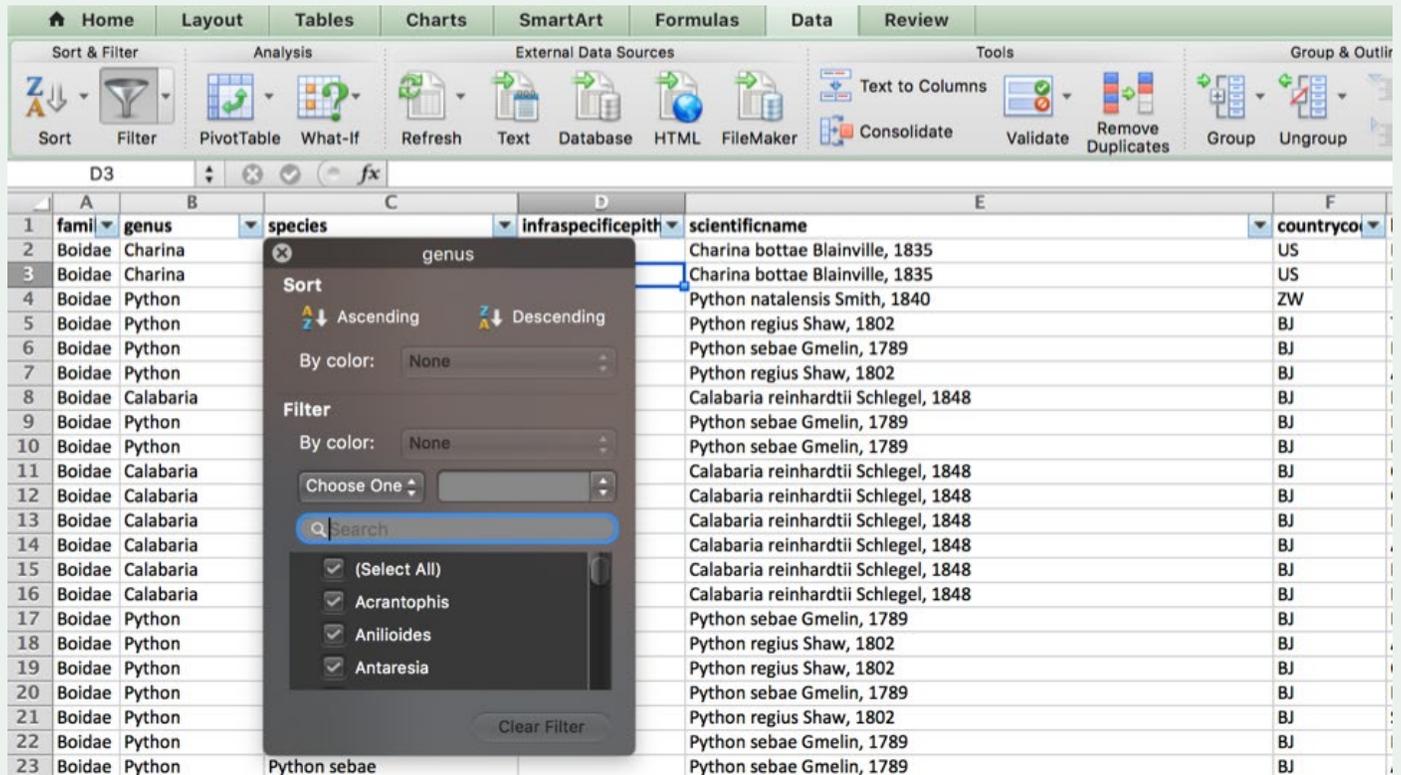
You can delete the other columns (i.e., anything other than columns A-L above). Be aware that sometimes GBIF changes the field names or format. Also, hereafter, the below instructions will assume you have your columns in the described order and position.

In your dataset, certain rows without any species or genus name may exist. You can delete such rows once you are sure you will not need them for later analyses. This can be done by adding a filter (see below), selecting the species with (blank) names, and then deleting those rows. You may also decide to keep the genus level record (e.g., to keep as many fossils as possible). If you decide to keep them, you can just replace the empty species cell with the value in the genus column. If the genus is unknown, then the information is most likely irrelevant, and you can just delete those rows.

### Tip: How to create a filter?

First, go to a cell in your table (e.g., cell A2 in the example), select all your table pressing command and A buttons, (Ctrl and A buttons for PC) and then click on Data > Filter.

The filter will allow you to show only certain rows without having to change the order of the table through sorting. To do this, go to the column you want to work with (e.g., column genus in the figure below) and press the triangle at the right side of cell. You will see a box with some options and the list of values in your column.



	fami	genus	species	infraspecific epith	scientificname	countrycode
1						
2	Boidae	Charina			Charina bottae Blainville, 1835	US
3	Boidae	Charina			Charina bottae Blainville, 1835	US
4	Boidae	Python			Python natalensis Smith, 1840	ZW
5	Boidae	Python			Python regius Shaw, 1802	BJ
6	Boidae	Python			Python sebae Gmelin, 1789	BJ
7	Boidae	Python			Python regius Shaw, 1802	BJ
8	Boidae	Calabaria			Calabaria reinhardtii Schlegel, 1848	BJ
9	Boidae	Python			Python sebae Gmelin, 1789	BJ
10	Boidae	Python			Python sebae Gmelin, 1789	BJ
11	Boidae	Calabaria			Calabaria reinhardtii Schlegel, 1848	BJ
12	Boidae	Calabaria			Calabaria reinhardtii Schlegel, 1848	BJ
13	Boidae	Calabaria			Calabaria reinhardtii Schlegel, 1848	BJ
14	Boidae	Calabaria			Calabaria reinhardtii Schlegel, 1848	BJ
15	Boidae	Calabaria			Calabaria reinhardtii Schlegel, 1848	BJ
16	Boidae	Calabaria			Calabaria reinhardtii Schlegel, 1848	BJ
17	Boidae	Python			Python sebae Gmelin, 1789	BJ
18	Boidae	Python			Python regius Shaw, 1802	BJ
19	Boidae	Python			Python regius Shaw, 1802	BJ
20	Boidae	Python			Python sebae Gmelin, 1789	BJ
21	Boidae	Python			Python regius Shaw, 1802	BJ
22	Boidae	Python			Python sebae Gmelin, 1789	BJ
23	Boidae	Python	Python sebae		Python sebae Gmelin, 1789	BJ

When the filter is applied, you can then select all the rows that are to be removed, by selecting the numbers in the left side of your screen. Then, once highlighted, delete all of these rows by selecting the option “delete” in the menu “edit”. Then, clear your filter to see the remaining data.

### Step 3. Finding synonyms

Synonyms are a common problem in taxonomy and occur when species thought to be different at first, are later found to be the same. Your GBIF dataset reports both the species as originally assigned to that particular record (column *scientificname*) and the current valid name (column *species*). The original scientific name column may include other information also, such as the subspecies (if any) or the authority information (the name and year the species was proposed). On the other hand, if the author of the occurrence (not of the species) was unable to identify the species, only the genus or even the family may be reported. To easily spot synonyms in your database use the F column and title it “Synonyms” and write this Excel formula below:

```
=(LEFT(E2,SEARCH(“”,E2,SEARCH(“”,E2)+1)-1))<>C2
```



as in

LEFT					
	A	B	C	D	E
1	family	genus	species	infraspeci	scientificname
2	Equidae	Equus	Equus niobrarenis		Equus alaskae (Winans, 1989)

This formula looks at the first two words (delimited by spaces) in the cell E2 (where the original scientific name is; column *scientificname*) and compares the result with the cell C2 (where the valid species name is; column *species*). If the scientific names in both the column *species* and column *scientificname* are different, you may have a synonym and the formula will return TRUE. If both of these columns match one another, the formula will return FALSE. Use the formula builder function in Excel to explore a bit more how it works, and remember, instructions inside brackets are solved first.

### Tip: Copying formulae into other cells

To copy the function to all the cells below it, move your cursor to the right bottom corner of the cell that contains the formula. The cursor will change in appearance to a black +. Double click on the bottom right corner and your formula will be copied to every cell below it. Excel will automatically copy the formula until it encounters an empty cell (either below, or in the cell adjacent to the left).

### Tip: Fast movements in the table

To move quickly through the table and identify breaks in the data that might prevent the complete copying of the formula, hold the Command key (Ctrl key in PC) and press the different directional arrows on your keyboard. That will allow you to jump through quickly, because it will look for changes between empty/non-empty cells. If you want to select all of these cells, then hold Shift and Command (Ctrl in PC) simultaneously, and then press the directional arrows.

Your table should like this:

LEFT					
	A	B	C	D	E
1	family	genus	species	infraspeci	scientificname
2	Equidae	Equus	Equus niobrarenis		Equus alaskae (Winans, 1989)
3	Equidae	Equus	Equus conversidens		Equus conversidens
4	Equidae				Equidae
5	Equidae				Equidae
6	Equidae	Equus	Equus conversidens		Equus conversidens Owen, 1869
7	Equidae				Equidae
8	Equidae	Equus	Equus simplicidens		Equus occidentalis (Leidy, 1865)

You will find different values that represent different situations:

- No synonym (Synonym value is FALSE), as in row 6 in picture above.
- Real synonym (Synonym value is TRUE), as in row 8 in picture above.
- The classification was done to genus or family level only (Synonym value is #VALUE!), as in rows 4, 5 and 7 in picture above.
- Incomplete information (Synonym value is #VALUE!), for instance, the authority information is not present in the column *scientificname* as in row 3 in picture above.

Using the filter in the column *Synonym*, select only the rows that are not FALSE; in other words select only empty cells, TRUE values, and errors such as #VALUE!. Once you decide a criterion is no longer needed (for instance, empty



cells are only caused by incomplete classification), you can update the filter by deselecting the particular value from the list.

In order to focus in on the actual synonyms, select only the TRUE values in the synonyms column filter. Now, you can copy the rows that can be observed in a new spreadsheet, and remove every column except for *family*, *genus*, *species*, *scientificname*, and *synonym*. Name the new spreadsheet “synonyms.xls”, remove the duplicates, and explore the synonyms in more detail. To remove duplicated rows, simultaneously hold Command (Ctrl in PC) and the letter “A” in your *synonym* spreadsheet and use the ribbon to perform the following: Data > Remove duplicates tool.

	A	B	C	D	E	F
1	family	genus	species	infraspecific	scientific	synonym
2	Equidae	Equus	Equus niobrarensis		Equus a	
3	Equidae	Equus	Equus conversidens		Equus c	
4	Equidae				Equidae	
5	Equidae				Equidae	
6	Equidae	Equus	Equus conversidens		Equus c	
7	Equidae				Equidae	
8	Equidae	Equus	Equus simplicidens		Equus o	
9	Equidae	Equus	Equus conversidens		Equus o	
10	Equidae				Equidae	
11	Equidae				Equidae	
12	Equidae	Equus	Equus conversidens		Equus c	
13	Equidae				Equidae	
14	Equidae	Equus	Equus conversidens		Equus c	
15	Equidae	Equus	Equus conversidens		Equus c	
16	Equidae	Equus	Equus conversidens		Equus c	
17	Equidae				Equidae	
18	Equidae	Equus	Equus conversidens		Equus conversidens Owen, 1869	FALSE

Save this spreadsheet—it will guide your exploration of the literature. For instance, when you are filling gaps using literature, you must use the new taxonomic name to be consistent with your table, but older papers may still be using synonyms of the species. This table can also help you to understand potential problems of classification that sometimes are explained in papers. Be aware that some of these problems are not fully solved by experts in the field.

#### Step 4. Getting the species, coordinates, and the fossils

Now you will move back to the original spreadsheet, because you need to recover the genus only information if available, the spatial information stored in the columns *decimallatitude* and *decimallongitude*, and you also need to distinguish fossils from extant specimens. First, we need to deactivate the Excel filter in order to ensure we are using our entire dataset. To do so, use your cursor to select any cell in your original spreadsheet (in order to make the spreadsheet “active”) and go to Data > Filter, and click on the filter button again to deactivate the filter.

#### Species and genus information:

Identifying species in the field or in the fossil records is hard, so you may find that sometimes genus information is the only information available. With your contrast taxon, browse through GBIF information as you did with your



focal taxon. There is no need to download the GBIF data for the contrast taxon. For your purposes, it is better to combine the genus and the species information in a single column. First, label the column *M* (that should be empty) as “SpeciesName”. Then, you will see if the information in the column *species* is empty. If it is empty, it means the species name is unknown, so we will try to get the genus name. The next formula will do these tasks for you:

```
=IF(ISBLANK(C2),B2,C2)
```

copy this formula to the column *M*.

Hint: Review the previous tips to quickly copy formulae.

Your table should look like this:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	family	genus	species	infraspecific	scientific	Synonym	country	locality	decimal	decimal	year	basisofrecord	SpeciesName	
2	Equidae	Equus	Equus caballus		Equus caballus	FALSE	US	Bosto	42.301	-71.1		PRESERVED_SPECIMEN	=IF(ISBLANK(C2),B2,C2)	
3	Equidae	Equus	Equus hemionus	Equus hemionus	Equus hemionus	FALSE	MN		43.106	106.9	2013	HUMAN_OBSERVATION	Equus hemionus	
4	Equidae	Equus			Equus quagga	TRUE	KE		-1.48	35.06	2014	HUMAN_OBSERVATION	Equus	
5	Equidae				Equus quagga	TRUE	KE		-1.298	34.85	2014	HUMAN_OBSERVATION	0	
6	Equidae	Equus	Equus quagga		Equus quagga	FALSE	KE		-0.197	37.01	2014	HUMAN_OBSERVATION	Equus quagga	
7	Equidae	Equus	Equus grevyi		Equus grevyi	FALSE	KE		0.263	36.91	2014	HUMAN_OBSERVATION	Equus grevyi	

See in row 4, that the genus information is retained. If there are no genus or species information as seen in row 5, Excel will return a “0” value.

*Fossils:*

Fossils represent the historic distribution of species and can also provide information about their distribution before humans started to have an impact on the Earth’s ecosystems. To distinguish between fossils and other type of occurrences, you will use the information in the column *basisofrecord*. First, you will label the column *N* (that should be empty) as “Fossil”. Then, you will look for those records that are explicitly identified as fossils (i.e., basis of record is “fossil\_specimen”). To do that, in the column *N*, use the formula:

```
=L2="FOSSIL_SPECIMEN"
```

Now, rows with TRUE values on column *N* represent fossils. Conversely, FALSE values in column *N* should represent non-fossils records, but GBIF is far from perfect and sometimes the basis of a record for an extinct species is recorded as “unknown”. So, if you have independent information that states a species went extinct before any direct observation of the specimen was possible (e.g., more than 30,000 years ago), then you must manually change the cell value of the column *Fossil* from FALSE to TRUE. Of course, you will need to do some background research on your focal taxon and identify species that are known to have gone extinct.

At the end, the right side of the table should look like:

	I	J	K	L	M	N	O	P
1	decimallatitude	decimallongitude	year	basisofrecord	SpeciesName	Fossil		
2	42.301069	-71.100285		PRESERVED_SPECIMEN	Equus caballus	=L2="FOSSIL_SPECIMEN"		
3	43.106121	106.935609	2013	HUMAN_OBSERVATION	Equus hemionus	FALSE		
4	-1.48011	35.064852	2014	HUMAN_OBSERVATION	Equus quagga	FALSE		
5	-1.298103	34.851533	2014	HUMAN_OBSERVATION	Equus quagga	FALSE		
6	-0.197487	37.011887	2014	HUMAN_OBSERVATION	Equus quagga	FALSE		
7	0.263024	36.908709	2014	HUMAN_OBSERVATION	Equus grevyi	FALSE		



### Coordinates:

You will summarize the information available into species present in squares of  $10^\circ \times 10^\circ$ . This will help to: 1) focus on global distribution patterns and 2) control for unequal sampling.

To summarize the information, you will round the latitude and longitude of each occurrence record to the closest multiple of 10. To do so, you can use some math and the function “round” that, as its name suggests, returns the closest integer. For instance, if the coordinate is 17.3, the latitude should be transformed into 20. To do so using the function round, you divide the number by 10, ( $17.3/10=1.73$ ), then you round the result ( $\text{round}(1.73)=2$ ) and you multiply the result again by 10 ( $2*10 = 20$ ). The Excel formula for this will be:

`=ROUND(I2/10,0)*10`

where “I2” refers to the second cell in the column *decimallatitude*. You can type this formula in the column O, and name it “Latitude”.

Do the same for longitude in the column P: “Longitude”. In this example the formula for this column will be:

`=ROUND(J2/10,0)*10`

Now, the right side of the table should look like:

	ROUND								
	I	J	K	L	M	N	O	P	Q
1	decimallatitude	decimallongitude	year	basisofrecord	SpeciesName	Fossil	Latitude	Longitude	
2	42.301069	-71.100285		PRESERVED_SPECIMEN	Equus caballus	FALSE	40	<code>=ROUND(J2/10,0)*10</code>	
3	43.106121	106.935609	2013	HUMAN_OBSERVATION	Equus hemionus	FALSE	40	110	
4	-1.48011	35.064852	2014	HUMAN_OBSERVATION	Equus quagga	FALSE	0	40	
5	-1.298103	34.851533	2014	HUMAN_OBSERVATION	Equus quagga	FALSE	0	30	
6	-0.197487	37.011887	2014	HUMAN_OBSERVATION	Equus quagga	FALSE	0	40	
7	0.263024	36.908709	2014	HUMAN_OBSERVATION	Equus grevyi	FALSE	0	40	

### Step 5. Summarizing the information

The previous step will link each of the species to a  $10^\circ \times 10^\circ$  square representing an area on the Earth. If more than one species occurs in a single location, then you will have more observations than are needed for your analysis, and is one of the first reasons to generate a table with only one record per species per square. When you plot the map (see below) for each species, you will be able to easily identify its global distribution. The second reason for generating a table with only one record per species is that, as you may recall, you will also need to review the global distribution of the number of species from this taxon present on Earth (so that you can discuss whether the taxon is primarily tropical or present only in the Americas, etc.).

To do so, copy the four columns we created in your most up-to-date table with occurrences (step 4) into a new spreadsheet, and paste it using Paste Special option (or clicking on the clipboard once pasted) and select the option “values only”.



	A	B	C	D	E	F	G
1	SpeciesName	Fossil	Latitude	Longitude			
2	#REF!	#REF!	#REF!	#REF!			
3	#REF!	#REF!	#REF!	#REF!			
4	#REF!	#REF!	#REF!	#REF!			
5	#REF!	#REF!	#REF!	#REF!			
6	#REF!	#REF!	#REF!	#REF!			
7	#REF!	#REF!	#REF!	#REF!			
8	#REF!	#REF!	#REF!	#REF!			
9	#REF!	#REF!	#REF!	#REF!			
10	#REF!	#REF!	#REF!	#REF!			
11	#REF!	#REF!	#REF!	#REF!			
12	#REF!	#REF!	#REF!	#REF!			
13	#REF!	#REF!	#REF!	#REF!			
14	#REF!	#REF!	#REF!	#REF!			
15	#REF!	#REF!	#REF!	#REF!			

- Keep Source Formatting
- Use Destination Theme
- Match Destination Formatting
- Values Only
- Values and Number Formatting
- Values and Source Formatting
- Keep Source Column Widths
- Formatting Only
- Link Cells

	A	B	C	D
1	SpeciesName	Fossil	Latitude	Longitude
2	Equus caballus	FALSE	40	10
3	Equus hemionus	FALSE	40	110
4	Equus quagga	FALSE	0	40
5	Equus quagga	FALSE	0	30
6	Equus quagga	FALSE	0	40
7	Equus grevyi	FALSE	0	40
8	Equus grevyi	FALSE	0	40
9	Equus quagga	FALSE	-20	30
10	Equus quagga	FALSE	-20	30
11	Equus quagga	FALSE	0	40
12	Equus quagga	FALSE	0	40
13	Equus quagga	FALSE	0	40
14	Equus quagga	FALSE	0	40
15	Equus caballus	FALSE	40	-10

And now it will look like

Now you can remove every column except for these: *SpeciesName*, *Fossil*, *Latitude*, and *Longitude*.

Select all the cells with values in your table and choose Data > Remove duplicates. You should get a table that looks like this:

	A	B	C	D
1	SpeciesName	Fossil	Latitude	Longitude
2	Equus caball	FALSE	40	-70
3	Equus hemio	FALSE	40	110
4	Equus quagg	FALSE	0	40
5	Equus quagg	FALSE	0	30
6	Equus quagg	FALSE	0	40
7	Equus grevyi	FALSE	0	40
8	Equus grevyi	FALSE	0	40
9	Equus quagg	FALSE	-20	30
10	Equus quagg	FALSE	-20	30
11	Equus quagg	FALSE	0	40
12	Equus quagg	FALSE	0	40
13	Equus quagg	FALSE	0	40
14	Equus quagg	FALSE	0	40
15	Equus caball	FALSE	40	-10
16	Equus caball	FALSE	40	-10
17	Equus caball	FALSE	40	-10

Remove Duplicates

- Select All
- Column A
- Column B
- Column C
- Column D

99296 duplicates found.  
1297 unique values will remain.

Remove Duplicates

Press “remove duplicates” and see the results. Copy the results to the spreadsheet, sppData, in the file *dataFrame.xlsx* provided by your instructor. Visually confirm that the order of the columns is consistent in both tables. If the column order is not consistent, reorganize the columns accordingly by cutting and pasting the columns into their proper positions.

As previously mentioned, some records do not have genus or species information. In those cases you will find a “0” in the SpeciesName column. You should delete those records as they provide little or no relevant information. To do that, add a filter and select with the filter the records with value “0” only. Then, press the row number selectors on the left and use the delete rows function in the edit menu, or right click and select the delete rows option. Finally,

	A	B	C	D	E
1	SpeciesName	Fossil	Latitude	Longitude	
39		TRUE	50	-100	
46	Cut	TRUE	40	-70	
52	Copy	TRUE	30	-100	
53	Paste	TRUE	40	-80	
56	Paste Special...	TRUE	30	-90	
58		TRUE	20	-100	
59		TRUE	10	-90	
61	Insert Row	TRUE	50	-120	
64	Delete Row	TRUE	30	-110	
73	Clear Contents	TRUE	0	40	
74		TRUE	10	40	
76	Format Cells...	TRUE	-30	20	
77	Row Height...	TRUE	50	0	
77	Hide	TRUE	0	-80	
78	Unhide	TRUE	-10	30	
78		TRUE	-30	30	
79		TRUE	0	30	
858	0	TRUE	50	10	
877	0	TRUE	-20	-70	
881	0	TRUE	-20	-60	
891	0	TRUE	10	-80	

remove the filter so that you can see all the records again.

**Step 6a. Understanding the data to map species distributions**

Now, let's take a look on the spreadsheets that represent the spatial distribution of oceans and species in the world. You can find the map in the spreadsheet *map* in *dataFrame.xlsx*, and the information used to create that map in the spreadsheet *data*. As you can see in both, the information in the spreadsheet *data* is a coarse resolution map of the world (10° latitude by 10° longitude grid), and it is linked to the spreadsheet *map* in the same file.

In the *data* spreadsheet, the column *A* (*tag*) indicates what that particular row refers to (species name, richness, ocean/non-ocean), column *B* (*Fossil*) indicates whether the report is a fossil or not, and columns *C* and *D* (*Longitude* and *Latitude*) represent the coordinates. For instance, a value "Ocean" in the column *tag* in the second row of the data spreadsheet means that everything in the second row in that spreadsheet contains information about the oceans. However, Column *E* (*Value*) is the actual value you will be storing, the other columns are often referred as identifiers. In the case of oceans, a number -1 indicates an ocean is there, while a number 0 represents a continent or island.

The ocean and continent information contained in *tag* and *value* columns in the data spreadsheet are needed to draw the map of the Earth. Excel will take the maximum value in the *value* column (Column *E*) that is present for any combination of latitude and longitude values, and color the cell accordingly. In other words, Excel will take into account if, for a given latitude and longitude, there is a value of -1 which denotes oceans, or a value of 0 which denotes continental areas, and will color your map accordingly. After you add your taxonomic data (instructions detailed immediately below), these colors will be further refined. Species presence or richness will have values that are larger or equal to one, therefore the maximum value for that coordinate will be larger than 0, and the color in that cell that will change accordingly. Now, take a look at the map and the data spreadsheets before you add your own data to make sure it looks correct.

**Step 6b. Building the data to map each species distribution**

Now we will add your species data in order to add taxonomic information to this background map of oceans and continents. To add your data, copy your results from the spreadsheet entitled *sppData* to the bottom of the table in the spreadsheet titled *data* that is in the file *dataFrame.xlsx*. Paste in your *sppData* beginning at row 705. Do not copy the row with the titles, just the data.

When copying the information to the spreadsheet *data*, be careful with the order of the columns, the columns must be in this order: *species*, *fossils*, *latitude*, *longitude*. In the *E* column (*value*) located within the spreadsheet *data*, add a number 1 for every single cell of the table that you just copied (do not add more cells than needed, only until the end of the table that you just copied here). This value of 1 that you just added to the column *E* corresponds to a presence of the species; this presence is then linked to a specific *latitude* and *longitude* which is reported in columns *C* (*latitude*) and *D* (*longitude*), respectively. This report of a taxon being present corresponds to either a current or historical distribution, as specified in the column *Fossil* (*B*). Your table should look like this:

	A	B	C	D	E
1	tag	Fossil	lat	long	number
702	ocean	NA	-90	160	0
703	ocean	NA	-90	170	0
704	ocean	NA	-90	180	0
705	Equus caballus	FALSE	40	-70	1
706	Equus hemionus	FALSE	40	110	1
707	Equus quagga	FALSE	0	40	1



### Step 7. Building the data to map the species richness distribution

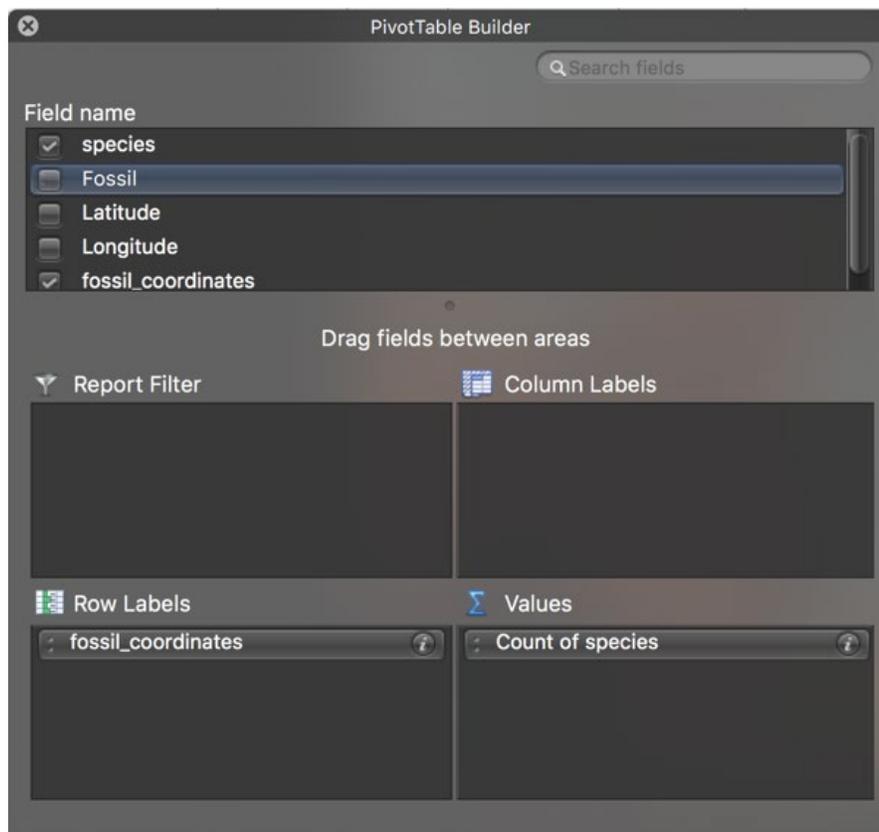
In this step, you determine the number of species in each one of the 10°x10° squares to find areas with a high density of species. You will also be able to distinguish between extinct and extant populations. To do this, you will work with the summarized information stored in the spreadsheet *sppData*.

First, let's go back for a minute to the spreadsheet *sppData*. Name column *E* (that should be empty) "fossil\_coordinates". Then copy and paste this formula and fill down for each row to combine the fossil, longitude, and latitude information in a single cell (be sure that the row number match the respective row):

=B2&"("&C2&","&D2&")"

	A	B	C	D	E	F	G
1	species	Fossil	Latitude	Longitude	fossil_coordinates		
2	Equus niobra	FALSE	20	-100	FALSE (20,-100)		
3	Equus conve	FALSE	20	-100	FALSE (20,-100)		
4	Equus simpli	FALSE	20	-100	FALSE (20,-100)		
5	Equus caloba	FALSE	20	-100	FALSE (20,-100)		

Select all the data in the spreadsheet *sppData* and build an automatic Pivot Table (Data > Pivot table, or Insert > Pivot table in newer versions of Microsoft Excel). In the Pivot Table window, first click on the word "fossil\_coordinates" located in the "Field name" section and drag this column into the "Row label" section. Then, click and drag the word "species", also located in the "Field name" section, into the "Values" section. In doing so, the word "species" will change to read, "Count of species". Do not enter anything into the "Column labels" field. This will give you the





number of extant and extinct species per 10°x10° grid. Now you will need to split the coordinate info again to get a table that you can append to the one in the data spreadsheet. To do this, copy the following formulas next to the result of the Pivot table in the cell indicated (we will assume that the Pivot table has the first column located as “A”, and the first row with values located at row number 5, as in the screenshot below):

Tag (cell D5): richness

Fossil (cell E5): =LEFT(A5,1)=""

Lat (cell F5): =MID(A5,SEARCH("(","",A5)+1,SEARCH(",",A5)-SEARCH("(","",A5)-1)+0

Long (cell G5): =MID(A5,SEARCH(",",A5)+1,SEARCH(")",A5)-SEARCH(",",A5)-1)+0

Value (cell H5): =B5

Now, fill in the formula for each column (D to H) for the remainder the Pivot table.

Your results should appear as in the figure below:

	A	B	C	D	E	F	G	H
1								
2								
3	Count of species							
4	Row Labels	Total		Tag	Fossil	Lat	Long	Value
5	FALSE (-10,-170)	1		richness	FALSE	-10	-170	1
6	FALSE (-10,-70)	1		richness	FALSE	-10	-70	1
7	FALSE (-10,10)	2		richness	FALSE	-10	10	2
8	FALSE (-10,130)	3		richness	FALSE	-10	130	3
9	FALSE (-10,140)	1		richness	FALSE	-10	140	1
10	FALSE (-10,20)	1		richness	FALSE	-10	20	1
11	FALSE (-10,30)	2		richness	FALSE	-10	30	2
12	FALSE (-10,40)	2		richness	FALSE	-10	40	2
13	FALSE (-20,-40)	1		richness	FALSE	-20	-40	1
14	FALSE (-20,-50)	1		richness	FALSE	-20	-50	1

Copy the results (columns D to H from the first row of data—in above example, row 5—until the last row with information in your table), and paste (using paste special, values only) this pivot table data at end (bottom) of the spreadsheet data.

### Step 8. Building the maps

At this point, you have all the information you will need in your spreadsheet data, and this should be linked to your spreadsheet *map* (be sure that there are no empty spaces in between the tables you just copied into the spreadsheet *data*). Now you can go to the map. The map is actually a pivot table linked to the spreadsheet data, using the ocean cells to give you an idea of the global distribution of your taxa, with latitude as rows and longitude as columns. Because each square represents a 10°x10° grid the continents appear a bit distorted. The sheet “map template” has a map of the Earth in a similar projection to give you an approximate idea the location of each square. You can copy the figure on top of your map and stretch it a bit, but it will never be perfect because Excel is not designed for these types of tasks.





To improve the readability of the map, the cells that belong to the map have been formatted using conditional coloring (Home > Conditional formatting). There are two conditions: gray for ocean and yellow-green for any value above 0. The yellow-green scale uses a continuous gradient between light yellow and dark green. The greater the value, the darker the green color. The formats are already set, so there is no need to change them. But just in case, here is the format used:

Now you can use the filter selectors (click icon in the cells B1 and B2; see above figure) to choose what you want to see (fossils? extant? one or more species, richness). In the map, -1 means ocean, 0 means continental areas without any record of the taxa, and any other value indicates that one or more species are present. Remember, to be able to see the map properly, you have to mark both i) the “ocean” in the tag filter, and ii) “NA” in the fossil filter.

### Step 9. Ranking the species distribution

The final thing you may wish to do is to use your database to rank your species in terms of the number of squares each species occupies. The more squares occupied, the larger a species’ range; in other words, the wider the species’ distribution. A species distributed everywhere in the world may be either: a) an invasive one (e.g., black rat); b) a species with a high dispersal capability (e.g., migratory birds); and/or c) a species possibly transported by humans (e.g., dogs). Conversely, the fewer squares a species occupies, the more endemic it is.

To get this ranking, you can go back to your spreadsheet *sppData* and create another Pivot Table in a new spreadsheet. This time, use the fossil column as “report filter”, species name as rows, and the combined coordinates as value. Now you can see the number of squares occupied by each species and can sort your Pivot Table to find the most widely distributed fossil or extant species, or the least. Here the first few rows for our example. Rename this spreadsheet as “taxaAbundance”.

	A	B
1	Fossil	FALSE
2		
3	Count of fossil_coordinates	
4	species	Total
5	Equus caballus	84
6	Equus asinus	57
7	Equus	38



### Step 10. Errors in the data

Cleaning a dataset is an important process, and it often requires several iterations. For instance, in this example, the most abundant species is *Equus caballus*, i.e., horses. However, horses and donkeys have been moved by human activities, so they are misleading when trying to understand geographic patterns. Therefore, it is better to map them, to keep track of the changes you are doing, and then remove them from the analysis. Since you have summarized all your data in the sheet *sppData* of the file *dataFrame.xlsx*, any update of your data (fixing a species name, updating the fossil status, adding or removing a line with data), have to be done in that table: the sheet *sppData* of the file *dataFrame.xlsx*. After you do it, you may have to repeat steps 6b-10.

#### Other ideas:

This is where your creativity begins. You can use these spreadsheets that you just created and the same tools to answer other questions that can help you in your assignment. But, this will depend on your specific taxon, as well as on the optional questions you choose to answer. Here are some ideas:

- How did the fossil record of species distribution change for your taxon over time? If you have a large number of fossils, can you create a timeline representative of the history of your taxon?
- How many countries has your taxa been found in? How can you infer “endemicity” using the number of countries and squares your taxon is present in?
- Can you infer endemicity using latitude and longitude? Do they give you the same answers?

### R Instructions (optional alternative to using Excel)

Below are instructions to process the information partially using R software. This section assumes you, the student, have some expertise in the usage of Excel and R, so the instructions are not as detailed as they are in the Excel version. Also, note that the logic of the steps are described in the Excel version of the instructions only, and not replicated here:

#### Step 1. Open the occurrence.txt file downloaded from GBIF

```
# read the data
oc <- read.delim("occurrences.txt", stringsAsFactors = F)
```

#### Step 2. Cleaning the data

```
# filtering columns
oc2 <- oc[,c("family", "genus", "species", "scientificname",
            "countrycode", "locality", "decimallatitude", "decimallongitude",
            "year", "basisofrecord")]
```

#### Step 3. Finding synonyms

```
# To find the synonyms, use a filter
synonyms <- oc2[(grepl("^([:alpha:]-)+ [[:alpha:]-]+", oc2$species) &
                (oc2$species !=
                 sub("^([[:alpha:]-)+ [[:alpha:]-]+.*$", "\\1", oc2$scientificname))),
               c("species", "scientificname")]
synonyms <- unique(synonyms)
names(synonyms) <- c("valid", "synonym")
write.csv(synonyms, "synonyms.csv", row.names=FALSE)
```

**Step 4. Getting the coordinates and the fossils**

```
# building the 10° squares
oc2$longitude <- round(oc2$decimallongitude/10)*10L
oc2$latitude <- round(oc2$decimallatitude/10)*10L
oc2$fossil <- oc2$basisofrecord == "FOSSIL_SPECIMEN"

# if there is no species information, fill with the genus
oc2$species[is.na(oc2$species) | oc2$species == ""] <-
  oc2$genus[is.na(oc2$species) | oc2$species == ""]
```

**Step 5. Summarizing the information**

```
# simplifying the table
sppData <- unique(oc2[,c("species", "fossil", "latitude", "longitude")])
sppData$q <- 1
# occurrences has now the information needed to paste
# on the "sppData" sheet of dataframe.xlsx.
# so you will export it as a csv file
write.csv(sppData, "sppData.csv", row.names=FALSE)
```

**Step 6. Building the data to map the species distribution**

```
# getting the data
data.spp <- sppData[,c("longitude", "latitude", "fossil", "species"),]
# removing rows without species
data.spp <- data.spp[!is.na(data.spp$species) & (data.spp$species != ""),]
# forcing species number to be 1, and changing the name of the species column
data.spp$number <- 1
names(data.spp)[4] <- "tag"
# exporting the data in csv format
write.csv(data.spp, "data.spp.csv", row.names=FALSE)
```

**Step 7. Building the data to map the species richness distribution**

```
# building the richness
richness <- aggregate(number ~ longitude + latitude + fossil, data.spp, FUN=sum)
richness$tag <- "richness"
richness <- richness[,c("longitude", "latitude", "fossil", "tag", "number")]
write.csv(richness, "data.richness.csv", row.names=FALSE)

# building the richness and spp data file
write.csv(rbind(data.spp, richness), "data.csv", row.names=FALSE)
```

**Step 8. Building the maps**

You can now import your file into Excel (because R will format your data in the .csv file format, which Excel recognizes). Therefore, you should be able to open your R output directly in Excel, by simply double clicking on the file in your file browser). At this point, if R output file (with the .csv file extension) opens seemingly in Excel please proceed to Step 8 of the Excel instructions presented above to build the maps, and then to Step 9 in R to build the next table.

**Step 9. Ranking the species distribution**

```
# count the number of squares where a species occurs.
spp.rank <- aggregate(number ~ fossil + tag, data.spp, FUN=length)
# rank them and
spp.rank <- spp.rank[order(spp.rank$fossil, -spp.rank$number),]
# export the file
write.csv(spp.rank, "rank.csv", row.names=FALSE)
```

Similar to the previous step, you can now import your file into Excel. After importing your data in Excel, paste it in an empty spreadsheet and analyze as suggested in Step 9 in Excel.

**Step 10. An error in the data**

If you find an error in your data, then you have to review the key table in your process, namely the table sppData. Because this is already in a csv file called "data.spp.csv" (see step 6), you can just open the file in Excel (or even with notepad), fix it, and read it again into R using:

```
data.spp <- read.csv("data.spp.csv", stringsAsFactors=FALSE)
```

After you read this file, go back and repeat steps 7 to 9, including updating your Excel file if needed.