

## DATA IS THE NEW BIOLOGY: AN INTRODUCTION TO AGGREGATED BIODIVERSITY DATA ON GBIF



### OBJECTIVES

Completing this module will enable you to:

- Explore biodiversity observations on GBIF
- Understand where different types of observations on GBIF come from
- Use GBIF to search occurrences by species, geography, and time
- Evaluate the utility of different data types for research questions

### INTRODUCTION

Biodiversity is defined as the variety of life on earth. Scientists study biodiversity at many different scales from the variation in the genes of a population to the diversity of species present in an ecosystem. It is important to understand that biodiversity has different facets and can thus be described in many ways, including taxonomic, morphological, genetic, ecological, and functional aspects. To address issues relevant to global biodiversity and its conservation (*i.e.*, climate change, zoonotic disease, invasive species, and extinction), new scientists like you may need to answer these questions by drawing on fields such as evolutionary biology, systematics, and ecology.

Advances in our ability to generate large amounts of biological data have transformed these fields. As a result, scientists like you must also increasingly learn new data science approaches. This can allow you to ask questions of larger and more complex types of data. In the life sciences in particular, there has been a rapid mobilization of data on the occurrences of species in space and time; where they occur, when they occurred there, and other properties. This information greatly increases our capacity to formulate conservation plans worldwide.

In this module, you will be introduced to a data platform called **GBIF** (Global Biodiversity Information Facility) as a gateway to explore these emerging, digital biodiversity data resources. You'll first execute searches of the data portal, narrowing results by species, geography, and time. You will be asked to think critically about the strengths and utility of these data resources and then encouraged to think beyond the obvious to how these data could be used to answer big questions in ecology, evolution, and conservation.

---

## ACTIVITY 1: EXPLORING OCCURRENCES ON GBIF

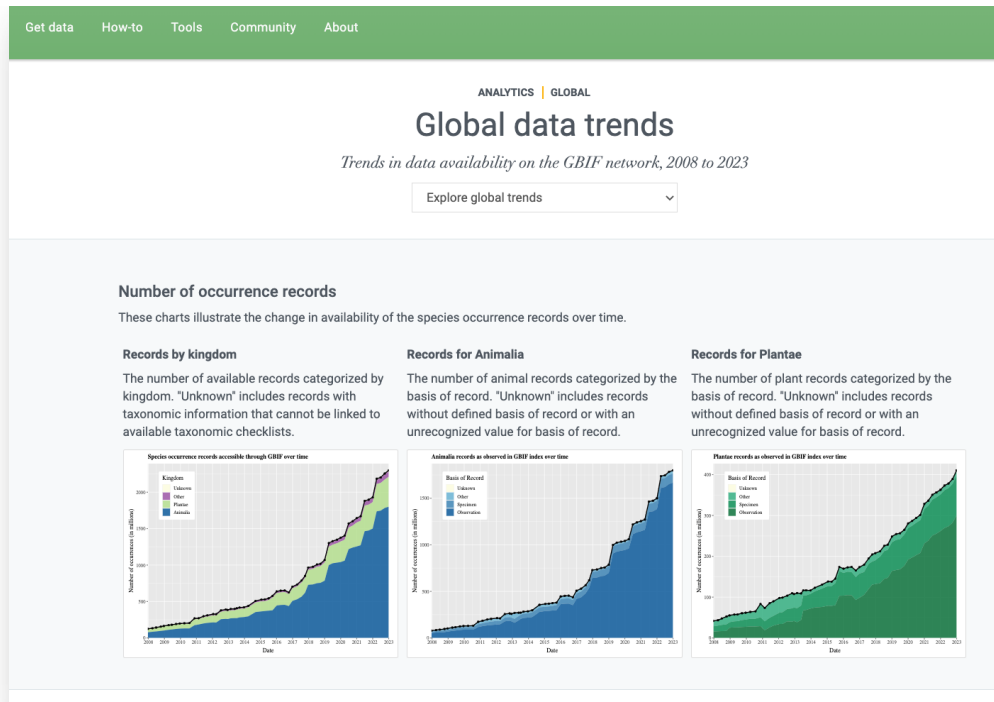
Descriptions of where and when a species was observed are called **occurrence data**. *Occurrences confirm the presence of an individual from a defined species at a specific place and time.* Occurrence data are fundamental to biodiversity research; they allow scientists to examine changes in distributions over time, perhaps in correlation with environmental factors, and to compare the distributions of different species.

There are several repositories that store occurrence data across the web, but today we will be focusing on one – GBIF (the **Global Biodiversity Information Facility**). Occurrence data on GBIF are derived from various sources, including collections of specimens in museums, field studies, citizen science observations, and others.

*A word of caution:* occurrence data, like all data types, have limitations. Some limitations might include uneven sampling effort, observation or collector bias, and incorrect taxonomic identifications. Part of biodiversity data literacy is considering these limitations and the suitability of the data for addressing the research question of interest. Researchers must also keep in mind that the **absence** of an occurrence record for a species at a location does not prove that it never existed there, only that an occurrence was never recorded there. Taking into consideration the ecology and behavior of your species of interest, as well different observation methods, will help you to correctly interpret and use occurrence data.

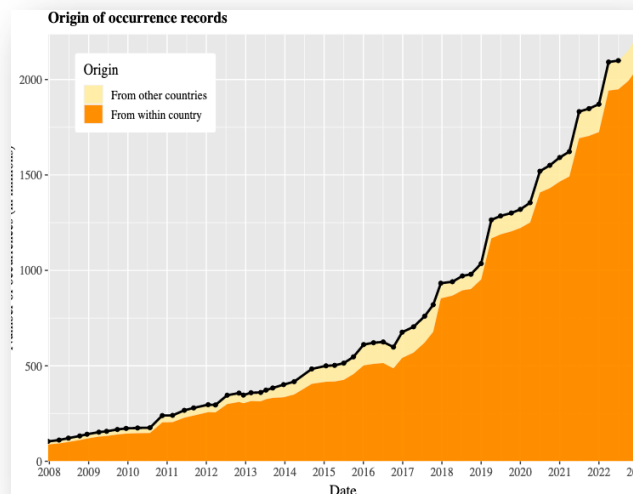
1. First, watch this [informational video](#) about GBIF.
2. Next, navigate to the **GBIF** landing page: <https://www.gbif.org/>. Statistics about the occurrence data in GBIF are shown here in near-real time. Answer the following questions based on what you see on this first screen.
  - a. How many **total biodiversity occurrences** are currently searchable through this portal?
  - b. How many **different datasets** do these records come from?
  - c. How many **different data providers/publishers** are there?
3. Go back to the top of the GBIF landing page, and expand the **Get data** tab. Here you will be given a list of options, pick the **Trends** option. This summarizes information about where all the diverse types of data on GBIF are coming from. Using the top row of graphs, answer the following questions.
  - a. What **taxonomic Kingdom** is currently represented the most in GBIF?
  - b. Approximately **how many plant and animal records** are contained in GBIF? (be sure to look at the Y-axis labels for units).

Figure 1. GBIF data trends as of [29 March 2023](#).



4. Finally, take a brief look at where data are coming from by scrolling all the way to the bottom section – **Data Sharing with Country of Origin**. Examine this plot to understand how each country is participating in data sharing. Do most occurrences come from *within the country of their origin*? Or from other countries?

Figure 2. Geopolitical origin for all GBIF data as of [29 March 2023](#).

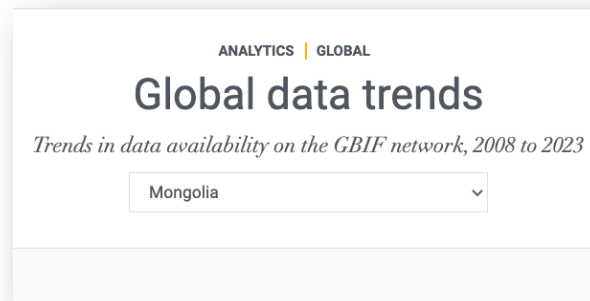


---

## ACTIVITY 2: COUNTRY-LEVEL SUMMARIES ON GBIF

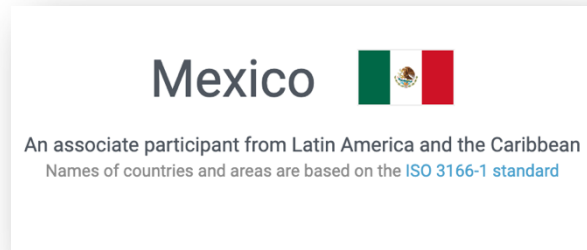
In this next brief activity, we will compare summaries of GBIF occurrences from two different countries of similar size (**MEXICO** and **MONGOLIA**). It is important to remember that, while GBIF is a global effort, not all countries may be able – or willing – to participate equally. The fact that some countries have higher biodiversity than others also affects the size and scope of occurrence records available, as does their history in conducting research.

1. Return to the GBIF landing page and navigate to the **Get data** tab, then select the **Trends** option. Let's do a more targeted search. We will target a predefined cultural/geographic region and explore species that have been observed there. You could use this for a research question, a social question, or even to explore biodiversity in a place you plan to visit.
2. In the pull-down menu at the top of the page, select the country of **Mongolia**. Scroll around on the map to remind yourself of where Mongolia is in the world. What countries surround it?



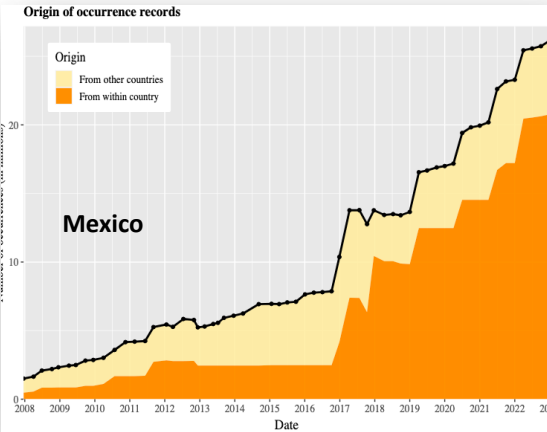
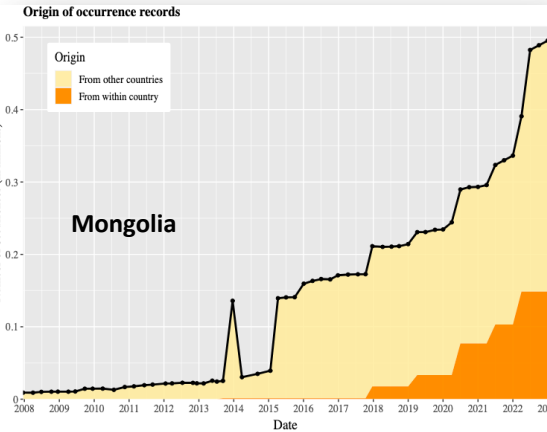
- a. How many **total occurrences** are there from Mongolia?
  - b. How many **total countries** have contributed these data?
  - c. Scroll further down to see the taxonomic breakdown of these occurrences.
    - i. How many of the occurrences are plants vs. animals?
    - ii. What other groups of organisms are on here that surprise you?
3. Next, scroll down and find the **Occurrences per Publishing Area** data box. This lists the places where most occurrences from Mongolia are coming from.
    - a. From **what five areas/countries** are most of the data coming from?
    - b. Take a moment to think about these patterns. What does they say about outside research interest in Mongolia? Do they say anything about countries that previously occupied Mongolia?

4. Go back to the pull-down menu at the top of the page and select the country of **Mexico**. Mexico is a similar size (in total area) to Mongolia, so we can use this search to ask how other factors such as total biodiversity and geopolitical location affect representation and participation on GBIF.



- a. How many **total observations** are there from Mexico on GBIF?
  - b. How many **total countries** have contributed these data?
  - c. Think back to how many occurrences there were from Mongolia. Approximately how many more (in terms of percentage) are there from Mexico?
5. As you did for Mongolia, scroll down and find the **Occurrences per Publishing Area** data box.
    - a. From **what five areas/countries** are most of the data coming from?
    - b. Take a moment to think about these patterns. Which of these reflect the research interests of outside countries, geographic proximity, or both? Which is most surprising to you?
  6. As a last task, scroll all the way to the bottom of the page and view the **Data Sharing with Country of Origin** plot. This is simply a visualization of how many occurrences come from within – versus outside of – the country you searched for. Compare the plots of Mexico and Mongolia below. What do you think this says about the history of research interest, the current scientific infrastructure, or both, within each country?

Figure 3. Geopolitical origin of occurrences from Mongolia (left) and Mexico (right) in GBIF as of 29 March 2023.



### ACTIVITY 3: SEARCHING BY TAXONOMY

Occurrence records on GBIF and other portals are anchored by three things: **taxonomic attributes** (e.g., a species name), **geographic attributes** (where the observation happened), and **temporal attributes** (when the observation happened). This activity will lead us through an occurrence search based on taxonomy, although we will remain somewhat geographically focused on Mongolia.

1. Pick a [Mongolian mammal species](#) that sounds interesting to you.
2. Next, search the GBIF portal for occurrences of this species. This search can be done by selecting the **Get data** tab, then the **Species** tab. Then type the Latin name of your chosen species. This is the default search function if you only desire to search based on taxonomy.
3. The results of your search should populate very quickly. In the search window, navigate to the **Metrics** tab. Here you can view an interactive summary of your search results. Focus on what types of observations your search yielded. The common basic types are:

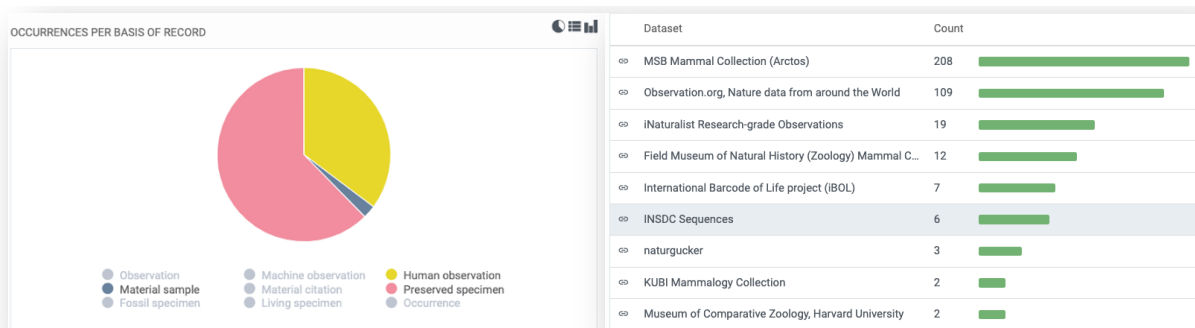
**Preserved specimen** (= specimen in a museum collection).

**Human observation** (= visualization by either a scientist or non-scientist).

**Material sample** (= information derived from an occurrence, like a DNA sequence).

4. Answer the following questions about your search results.
  - a. **How many records** of your chosen species were found in GBIF?
  - b. What is the most common **Basis of Record** of your occurrences? Name the top three.
  - c. What is the most common **Dataset/Data Provider**? Look at the top three. What Basis of Record Category do the datasets correspond to (iNaturalist, museum collections, etc.)?
  - d. Thinking about museum collections only, what are the top three museums your occurrence records are coming from?

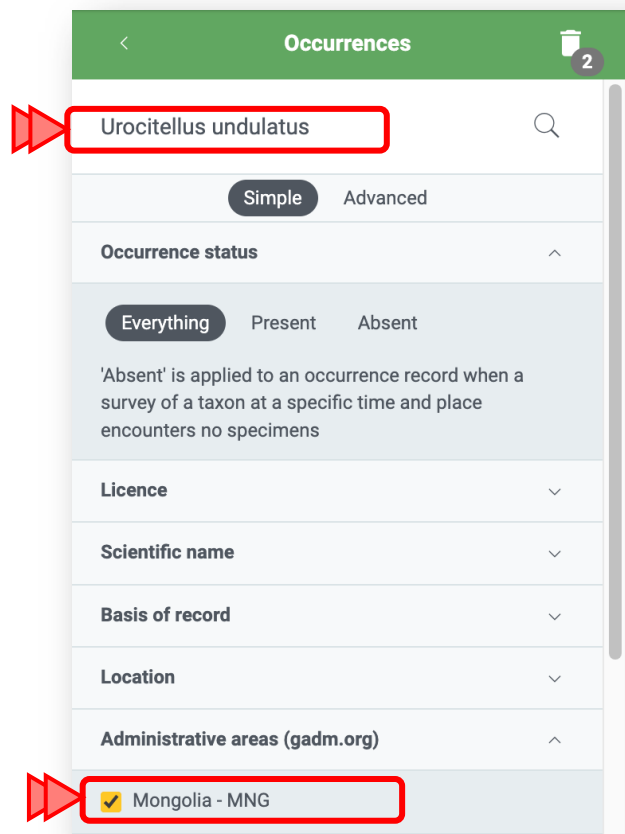
Figure 4. Basis of record, and most common data providers, for the long-tailed ground squirrel in GBIF as of 29 March 2023.



## ACTIVITY 4: SEARCHING BY TAXONOMY, GEOGRAPHY, AND TIME

As covered above, occurrence records have **taxonomic attributes** (e.g., a species name), **geographic attributes** (where the observation happened), and **temporal attributes** (when the observation happened). In this final activity, we will conduct a search parameterized by each of these things. This is the beginning step of more in-depth data exploration and can help you develop biologically relevant questions.

1. Go back to the [Mongolian mammal species](#) list and locate the long-tailed ground squirrel (*Urocitellus undulatus*). This species inhabits meadows and steppes across much of Siberian and interior Asia.
2. Select the **Get data** tab, then the **Occurrences** tab. This is the advanced occurrence search function in GBIF, and we are using it to search on taxonomy, geography (here, the geopolitical unit of country), and year. Search for records of the species in Mongolia, as below:



The screenshot shows the GBIF Occurrences search interface. The search term "Urocitellus undulatus" is entered in the search bar. The search mode is set to "Simple". The Occurrence status is set to "Everything". The location is set to "Mongolia - MNG".

Urocitellus undulatus

Simple Advanced

Occurrence status

Everything Present Absent

'Absent' is applied to an occurrence record when a survey of a taxon at a specific time and place encounters no specimens

Licence

Scientific name

Basis of record

Location

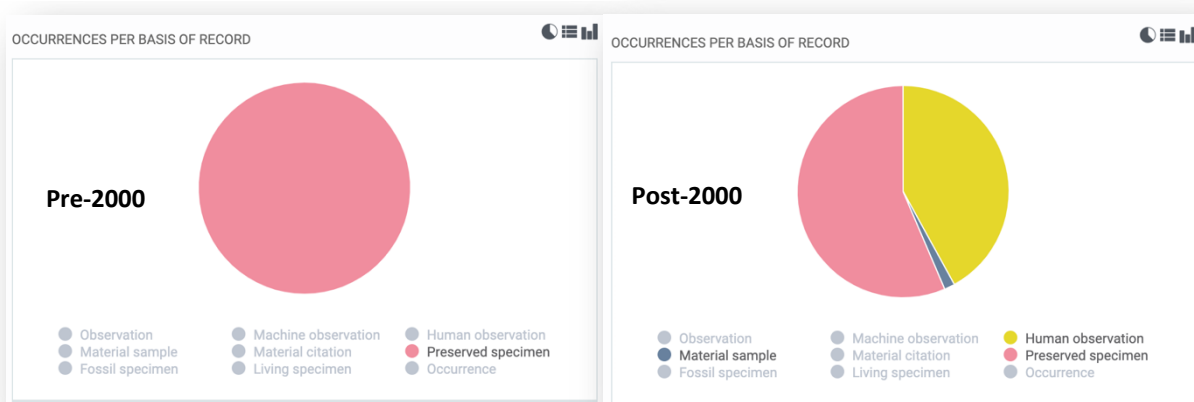
Administrative areas (gadm.org)

Mongolia - MNG



3. In the search results window, navigate again to the **Metrics** tab. View the different summaries of your search results, and answer the following questions.
  - a. **How many records** of this species in Mongolia were found in GBIF?
  - b. What are the two most common **Basis of Record** of your occurrences?
  - c. What are the top three **Datasets/Data Providers**? What Basis of Record Category do the datasets correspond to (citizen science, museum collections, etc.)?
  
4. Our last task is to include the element of **time**. To do this, we will conduct **two more, separate searches** that limit occurrences by the year of observation.
  
5. First, filter your search results for *U. undulatus* in Mongolia to everything **PRIOR TO 2000**. To do this, find the **Year** search box and use the slider tool to select all years before 2000.
  - a. **How many records** were left?
  - b. What is the most common **Basis of Record**?
  - c. Click over to the **Gallery** tab in the search results window. Do you see any occurrences that have photographs of this species?
  
6. Next, change to filter your search results for *U. undulatus* to everything **AFTER 2000**.
  - a. **How many records** exist for the 21<sup>st</sup> century?
  - b. What is the most common **Basis of Record** for these recent records?
  - c. Click over to the **Gallery** tab in the search results window. You should see numerous (cute!) photographs of this species, reflecting the influx of citizen science records as these platforms have become more widely used. Two of the largest platforms are *iNaturalist.org* and *Observation.org*. Are these providers listed in the **Metrics** tab?

Figure 5. Basis of record for the long-tailed ground squirrel in Mongolia before (left) and after (right) year 2000.



**Great work!** You've learned how to explore biodiversity data and occurrence properties on GBIF. Next week we will start out class by looking more closely at the two main data types/providers we have encountered: museum collections (on Arctos) and citizen science observations (on iNaturalist).

---

## POST-ACTIVITY QUESTIONS

### Section 1: Multiple Choice (select only one answer!)

1. Each of the following describes GBIF **except which one**?
  - a. A social network of people interested in nature
  - b. A place to view nature observations from citizen scientists
  - c. A place to explore museum collection holdings from across the world
  - d. A platform to explore biodiversity occurrences
  
2. Where do occurrences on GBIF come from?
  - a. Field surveys
  - b. Citizen scientists
  - c. Museum collections
  - d. Government agencies
  - e. All of the above
  
3. Which of the following would NOT be considered a primary biodiversity **occurrence**?
  - a. A record of a species at a place and time
  - b. An undated and unplaced picture of a zebra from your great-grandfather's trip to Africa
  - c. A Galapagos bird specimen with time and place data, donated to a museum by Charles Darwin
  - d. A digital image of a cricket species uploaded to iNaturalist, with a date and latitude/longitude
  
4. What can you use GBIF for?
  - a. Searching for species in your geographic area
  - b. Visualizing the approximate distribution of a single species of interest
  - c. Uploading a picture to receive help identifying species you encounter
  - d. All of the above
  - e. A and B
  
5. Users can filter a search on GBIF in each of the following ways, **except which one**?
  - a. By taxonomy (species identity)
  - b. By place and time
  - c. By data provider
  - d. None of the above, GBIF only provides a list of occurrences

