|  |  |
| --- | --- |
| **An Introduction to Biodiversity Databases and Specimen Images** |  |

\*This lab was modified from the BLUE network module by Monfils, A., Linton, D., Ellwood, L., Phillips, M. (2019). [Data is the New Science](http://dx.doi.org/10.25334/Q4RR0R).  [Biodiversity Literacy in Undergraduate Education](https://qubeshub.org/groups/blue_data), QUBES Educational Resources. [doi:10.25334/Q4RR0R](http://dx.doi.org/10.25334/Q4RR0R)

## Pre-Survey

1. Have you visited a natural history museum or aquarium before?
2. Do you feel it is important to preserve biological specimens in a natural history museum?
3. Every biological specimen in a museum (examples: a squirrel, a mushroom, a flowering plant) has information about that specimen recorded on a label. What information do you think is recorded on that label? Try to think of at least two pieces of information, but list as many as you like.
4. The Long Beach State Herbarium (preserved plant museum) (LOB) contains 18,000 specimens. In the United States there are 659 herbaria with an estimated 76.5 million specimens. What percentage of the United States' specimens do you predict are accessible to you in an online database? (could be as photos, could be specimen label info only).

## Objectives

Students completing this module will be able to:

* Access data from biodiversity digital data repositories CCH2, iDigBio, GBIF, and iNaturalist
* Evaluate the research utility of occurrence data derived from different sources
* Create and interpret a graph
* Use geo-spatial data to inform biological thinking
* Access specimen image data for morphometrics study

**Introduction**

There is a changing landscape for those joining the 21st century workforce. Rapid advances in data research and technology are transforming how we conduct science. The volume and variety of data being generated, the increased accessibility of data for aggregation, the improved discoverability of data, and the increasingly collaborative and interdisciplinary nature of scientific research are driving the need for new skill sets.

The biodiversity sciences, for example, have experienced a rapid mobilization of data that has increased our capacity to investigate large-scale issues. **Biodiversity** is 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. Biodiversity can be described in many ways, including taxonomic, morphological, genetic, ecological, and functional. As we look around our world, all the many forms of living organisms we see are the product of over 3.5 billion years of evolution in the form, function, and processes of life on earth. In order to address scientific issues of critical national and global importance, such as climate change, zoonotic disease, resource management, invasive species, and biodiversity loss, the 21st century biodiversity scientist must be fluent in integrative fields spanning evolutionary biology, systematics, ecology, geology, genetics, biochemistry, and environmental science and possess the quantitative, computational, and data skills to conduct research using large and complex datasets.

In this module, you will be introduced to some emerging biodiversity data resources. 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 science questions. You will then use some of these databases to find images of plant specimens that we will use for our morphometric projects.

**Activity 1: Examine Specimen-Based Occurrence Data: What is the distribution of Marah macrocarpa?**

The data we collect about the where and when a species is found is called **occurrence data.** They are information about the presence of an individual from a particular species at a specific place and time. Information from occurrence records has been used extensively in research, allowing scientists to examine changes in distributions of species over time, perhaps in correlation with specific environmental factors, and to compare the distributions of different species. The **evidence** that a species is present can simply be a record of someone observing the species is present or the evidence can be the collection of specimens, which are then preserved in a natural history collection.

***Specimen-based data*** are based on archived biological specimens housed in a natural history collection. Scientists, naturalists, explorers, and students have been collecting and preserving specimens for hundreds of years. Today, we can interact directly with the specimens collected by Darwin, Captain Cook, Lewis & Clark, Teddy Roosevelt, and Linnaeus. Preserved with the specimens is information about the organism (e.g., collector, collection date, location, habitat, images, nearby species, phenology). Specimens can be further examined to verify the information (Does it have the correct scientific name applied to it? Was it really collected in that location?) or to yield additional information as scientists build on previous research or identify new questions to investigate. Preserved specimens from natural history collections are a treasure trove of data and information and have been used to study a variety of topics, such as extracting DNA to study evolutionary relationships, assessing pesticide use, host-parasite evolution, and zoonotic disease transmission, etc. In 2018, the National Science Foundation funded California museums and universities to begin a massive effort to produce images of herbarium (plant) specimens and make them available to the public in a digital database. The images are aggregated in searchable portals that make them accessible to researchers and the public. The database contains information on individual specimens, images of the specimens, and associated data.

1. Go to the website for the **Consortium of California Herbaria2 (CCH2)**: <http://cch2.org/portal/> . This database contains plants, lichens, and algae.
2. Then go to Search Collections, leave all Collections selected, and click “Search.” Type the name “***Marah macrocarpa***” into the query bar and search by List Display.

 

Figure 1. *Marah macrocarpa* or ‘chilicothe’ is a wild cucumber, native to California. The plant has a vine habit and lobed leaves with palmate venation. The prickly capsule fruit develops from an inferior ovary and, unlike our garden cucumber, the fruit is inedible. Photos by A.E. Fisher

* 1. How many records of your species were found?
  2. How many of the specimens are from the California State University, Long Beach Herbarium?

1. Click on the maps tab and Display coordinates in Google Map. Do any of the dots seem to be far away from the other dots? These are outliers in the distribution. They may be accurate records or they may be a mistake in the database. Click on 2 or 3 outlier dots and check if their written location matches where they are located on the map.
   1. Use a colored pencil to outline the distribution of your species in CCH2 on the map below or insert a screenshot of the CCH2 distribution here.



credit: pixaby.com [CC0 Creative Commons License]

1. Go back to Search Criteria and keep the scientific name in the query bar. Scroll down and under Specimen Criteria, click Limit to Specimens with Images. Then Search as List Display.
   1. How many records of your species were found with images?
   2. What percentage of occurrence records in the CCH2 database have an image for this species?
2. The images can be opened and downloaded as large files. Choose an image to click on and “Open Large File.” Once the file downloads, click on the image to zoom in and see the fine scale resolution.
3. Notice that the herbarium specimens are photographed with a scale bar on the image that allows us to measure the length of plant structures in the image.
4. Next, go to the **iDigBio** website: https://www.idigbio.org . This database contains records for all kinds of biological specimens, including plants, animals, and fungi.
5. How many specimen records are currently available through the iDigBio portal?
6. Use the same species as above (*Marah macrocarpa*) and search the iDigBio portal for occurrence records.
7. How many records of your species were found in iDigBio? Is this more or less than the number of *Marah macrocarpa* records in CCH2?
8. Use a different color to outline the distribution of your species on the map you made from CCH2 data or insert a screenshot of the iDigBio distribution here.
9. In iDigBio, what is the date of collection for the oldest preserved specimen of *Marah macrocarpa*? Hint: Click on the heading “Date Collected” to re-order the spreadsheet using that column.
10. Click on the “must have media” button. Once the dataset refreshes, click on the Media tab near the bottom of the page. What types of media are available for researchers to use?
11. Next, go to **iNaturalist***.*<https://www.inaturalist.org/> This database contains ***observation-based data*,** or an individual biological organism observed by scientists, naturalists, and citizen scientists. Observation data is not necessarily linked directly to a physical specimen in a natural history collection. While some observations may be associated with just a location and date, most observations are accompanied by a photo and detailed information (e.g., collection methods, environmental conditions, geographic location, associated species, weather, behavior, abundance, phenology). Similar to specimen-based data these data can provide a wealth of information on biodiversity and are now included in some online databases accessible to researchers, educators, and the public. Many people around the world use the iNaturalist app to identify and observe organisms. Try it out!
12. Search for the same species, *Marah macrocarpa,* in iNaturalist.

* 1. How many records of your species were found in iNaturalist? Is this more or less than the number of *Marah macrocarpa* records in CCH2 or in iDigBio?
  2. If there are any outliers, click on several and check if their image matches what the species looks like. Botanists have described *Marah macrocarpa* as a California and Baja California species. There are 55 other species of wild cucumbers known from the United States and Canada. Flora of North America Cucurbitaceae: [Illustration](http://www.efloras.org/object_page.aspx?object_id=128298&flora_id=1)
  3. Notice that the photographic evidence for iNaturalist observations only rarely have an object for scale. This means we have no way to measure the true size of the plants in the image.
  4. Do you think data from iNaturalist is as reliable as data from a museum or university natural history collection? Why?
  5. Add any additional valid points for *Marah macrocarpa* from iNaturalist to the distribution map above or insert a screenshot of the iNaturalist distribution here.

1. Finally, go to the Global Biodiversity Information Facility, **GBIF**.org. This database includes both specimen and observation-based occurrence records. CCH2 and iNaturalist send their observations to GBIF, but it also includes other sources, including collections from many other museums from around the world. At this time (2019) GBIF is likely the most inclusive source of occurrence data for most biological organisms.
2. Search the GBIF portal for occurrences of *Marah macrocarpa*. Refer to the GBIF User Guide as needed to help navigate the portal.
   1. How many total occurrences of the species were found?
   2. Click on the green box with the number of occurrences to display the occurrences table. Then in the left side bar click on the Basis of Record. How many of the occurrences are Preserved Specimens?
   3. What are some of the other basis of records categories?
   4. Select Preserved Specimens and, once the occurrences table loads go to the Gallery. Choose an image and zoom in on it to see the resolution. These images can also be downloaded.
3. Go back to the search results and click on the Map tab. Outline the distribution of your species on the map or insert a screenshot of the GBIF distribution here.
   1. How do the distributions from CCH2, iDigBio, iNaturalist, and GBIF compare to each other?
   2. Which data source, CCH2, iDigBio, iNaturalist, or GBIF, should you use to generate the best estimate of your species’ distribution? Justify your answer by discussing the number of occurrence available, types of records (observation or specimen-based), accuracy of identification, and the ability to verify records.

**Activity 1 Summary**: Occurrence data, like all data types, have limitations. Some limitations of occurrence data include uneven sampling, observation or collector bias, and incorrect identification. Part of data literacy is considering the limitations of data and the suitability of data for addressing the research question of interest. For example, when using occurrence data, the researcher needs to keep in mind that the absence of an occurrence record for a species at a location does not prove that it has never existed there, only that an occurrence was never recorded there. It is also important to take into consideration the biology of your species of interest. Considering the biology and ecology of your organism or ecosystem of interest, combined with an understanding of the sampling methods for gathering data and the complexity of the data records, are all critical aspects of working with occurrence records.

**Activity 2: Choosing two species to compare for your Geomorphometrics project**

1. Return to CCH2.org and Browse Images <- Browse by Genus. You may also use the plant reference books we have in the classroom. Choose a genus with at least two species listed. Then search for specimens with images in CCH2.
2. Check that your two species meet the following criteria:
   1. Are there at least 5 unique specimen images available for each species? Check CCH2, GBIF, and possibly iNaturalist (only use if there is a scale bar in the image).
   2. Do the specimen images show flat leaves and do they have flowers or other reproductive structures?
   3. Is there information on the label about where and when the specimen was collected?
   4. Are the images of the correct species? Sometimes the wrong image is linked to the database occurrence. Ask the instructor to check them over if you're not sure.
   5. You will need to make measurements of two structures on these specimen images. We will work on identifying the two organs your group will study in another lab.
   6. What are the scientific names of the two species you might compare for your geomorphometrics project?
   7. What are two structures, one vegetative and one reproductive, you might measure and compare between these species? Students in the past have often chosen leaves and petals, but it depends on which species you choose.

Every student will find two species and at least two organs to compare. Then we will form our project groups and pitch our ideas to each other.