|  |  |
| --- | --- |
| **The Power of Data Standards** | **A close up of a sign  Description automatically generated** |

## Objectives

Upon completion of this module, each student should be able to:

* Explain why standardizing descriptions and organization of data are important.
* Explain problems or limitations of using data when not standardized or organized.
* Strategize about the best ways to organize data to answer a question.
* Apply their knowledge of data quality standards when collecting/creating their own data.

|  |
| --- |
| “*After the lapse of many years, possibly a century, the student of the future will have access to the original record of faunal conditions in California*”.  - Joseph Grinnell |

**Introduction**

Joseph Grinnell (1877-1939) was an ecologist and the first director of the Museum of Vertebrate Zoology at Berkeley. He is a great example of how forethought and careful record-keeping can have lasting impact in science. Grinnell took regular expeditions into the field and collected many specimens, primarily in the western US. He took detailed notes in the field on both abiotic conditions and biotic communities during these trips. When he added specimens to the museum collection, he was able to meticulously catalog information on when and where the animal was collected. Here’s what he wrote back in 1910:

*"It will be observed, then, that our efforts are not merely to accumulate as great a mass of animal remains as possible. On the contrary, we are expending even more time than would be required for the collection of the specimens alone, in rendering what we do obtain as permanently valuable as we know how, to the ecologist as well as to the systematist. It is quite probable that the facts of distribution, life history, and economic status may finally prove to be of more far-reaching value than whatever information is obtainable exclusively from the specimens themselves."*

From: "The Methods and Uses of a Research Museum" Popular Science Monthly 77: 163–169. (Adapted from Museum of Vertebrate Zoology, 2015).

**Activity 1: Sample Data Record**

Recent efforts to digitize museum collections mean the specimens collected years ago can be accessed by researchers around the world. This means we can look at changes in morphology, range distribution, or habitat affinity over time. However, in order for these data to be useful for analyzing change over time, there needs to be some consistency in the way they are collected and reported. In the following activities, we will demonstrate the importance of data standardization.

Throughout this module, we will be using a case study of the Alpine Chipmunk (*Tamias alpinus)*. Your role is as a researcher completing a follow-up study on the range distribution and sex ratio of the Alpine Chipmunk. Researchers suspect this animal has been impacted by changing climate in recent years.

What information would you hope to be able to access from historical data?

*Procedure*

1. Assume you are in the field collecting data on the range distribution and sex ratio of the Alpine Chipmunk in California and you have caught a chipmunk in a trap. Before you release the chipmunk, you will need to record data. Create a sample entry that you might record in your field notes, including both the data fields and data values.
2. Consider the different ways in which a researcher studying Alpine Chipmunks could record the date, location, or sex of chipmunks caught in traps. Look at the sample field notes you developed and compare the format of your entries to your partner’s field notes. How are they similar? How are they different?

**Activity 2: Data Exploration**

To investigate some of the potential pitfalls of unstandardized data, we are going to explore some historical data in digitized natural history collections. For a sample dataset we will look at specimens that Joseph Grinnell collected during his lifetime.

*Procedure*

1. Open the iDigBio portal at <https://www.idigbio.org/portal/search>.
2. In the search field “Collected By” search for “Grinnell”. You may need to “Add a Field” under Filters because this is not one of the default search fields.
   1. How many total records are returned?
   2. Look at the map. Are there any data points that don’t seem to make sense? If so, where are they and why do they not make sense?
3. In the records section, click on “Columns” and check the box to add the “Collected by” column to the results.
   1. Were all of the specimens collected by Joseph Grinnell?
   2. Is the format for the names of collectors the same for all of the collectors listed? If not, what different variations are present?
4. Look at the “Countries” listed in the results.
   1. List the first four *distinct* entries.
   2. What inconsistencies are present in this data field and what problems might this present for researchers?
5. Look at the ‘Date Collected’ column.
   1. Are the data written in the same format?
   2. If you were interested in downloading this dataset to use for your own project, why might this cause a problem?
6. Look at other columns, such as ‘Scientific Name’ and ‘Family’. (Note: if these are not displayed, you can click on the “Columns” box on the right side to add).
   1. Are these data standardized?
   2. Why may there be different ways of entering this information?

**Activity 3: Data Standards**

Standards make it easier to collect, share, assimilate, and use data by establishing guidelines on how data are represented, ensuring the data are in a format that can be easily used, and systematizing the meaning of the data. "Standards provide data integrity, accuracy, and consistency, clarify ambiguous meanings, minimize redundant data, and document business rules" (USGS, 2019). When standards are maintained, data can more readily be integrated across projects and researchers.

In the above exploration, you identified differences in the formats of many data fields. An example from USGS (2019) that relates to data ambiguity and the need for standards demonstrates some of the challenges with how dates are recorded. Note the variation in these formats for dates, keeping in mind that people in some countries write the month first, while others record the day of the month first.

* April 2, 1974
* 04-02-74
* 04/02/1974
* 4/2/74
* 19740402
* 04021974 - is this April 2 or February 4?
* 2 April 1974

To integrate datasets from databases using these different formats for date, a researcher would first have to convert the dates into a common format. Utilizing data standards would establish a consistent structure, making integration and use of multiple datasets much easier.

Managers of large sets of biodiversity data collect, manage, and maintain a wide variety of information in repositories such as natural history collections at museums, herbaria, monitoring programs, and citizen science projects. Data deposited in these collections varies in format and detail. In order to compile data simply and efficiently, the Darwin Core Standard was established to provide a mechanism for integrating biodiversity datasets from various sources (GBIF.org 2019).

*Procedure*

1. To get a glimpse of how Darwin Core works, go to<https://dwc.tdwg.org/terms/> to look at some of the formats for data standards in Darwin Core.
2. Choose “Event” in the column on the right and then choose “eventDate” in the tab at the top.
   1. Do the data from your search in iDigBio match the Darwin Core recommendations for date standards?
3. Look at some of the other fields. Review your sample field notes on the chipmunk data from Activity 1.
   1. What is an example of a data field that you would need to consider in order to record your data using a Darwin Core standard?

**Activity 4: Refining a Search**

Okay, let’s get back to looking for Joseph Grinnell’s specimens.

*Procedure*

1. Use the following search terms to refine your search:

Collected By: Grinnell

Class: Mammalia

Date Collected: start: 1877-01-01 end 1939-12-30

(these are the years Joseph Grinnell was alive, although it is not likely he was making collections the year he was born!)

1. How many specimens do you have now?

Let’s focus in even more. Imagine we want to look at historical changes in the occurrence of the alpine chipmunk, *Tamias alpinus*.

1. Clear your search (reset at the top of the screen) and search for *Tamias alpinus* (scientific name).
   1. How many records did the search identify?

Let’s refine the search even further. During the time period 1914-1920, Joseph Grinnell and colleagues surveyed small mammals in California, including Yosemite National Park (YNP). The “Yosemite Transect” documented 4354 specimens and yielded over 3000 pages of field notes and approximately 700 photographs (Moritz et al. 2008).

1. Narrow your search for *Tamias alpinus* to the years from 1914-01-01 to 1920-12-31.
2. Use the rectangle or circle bounding boxes on the left side of the map to zoom in so to include only the specimens recorded from YNP (north of Fresno, CA).
3. Record the distribution of *T. alpinus* on the map below.

A screenshot of a computer screen

Description automatically generated

**Activity 5: Comparing Historical and Modern Data**

Thoroughly and meticulously recording data can be beneficial for scientists who want to study the impact of climate change on the distribution of species. The data recorded by Grinnell and colleagues provide a benchmark for comparing current distributions of small mammals like *T. alpinus*. From 2003-2006, a group of scientists resampled the Yosemite Transect to determine if climate change over the past century had impacted the small mammal community in YNP. They were able to compare historical and modern surveys thanks to the precise record-keeping of species and sampling sites by Grinnell and colleagues. Since YNP has been protected since 1890 and has experienced fewer human-caused changes in land use, they were able to correlate changes in small mammal distribution to climate change (Moritz et al. 2008).

*Procedure*

1. Clear your search and see how many records you can find for *Tamias alpinus* in the time period 2003-01-01 to 2006-12-31.
2. Zoom in on the map to include only the specimens recorded from YNP.
3. Record the modern distribution of *T. alpinus* on the map above in a different color and add a key to indicate which color refers to which time period.
   1. What do you notice about the distribution of current records of *T. alpinus* compared to the records recorded by Joseph Grinnell and colleagues?
4. Review the figure below from Moritz et al. (2008). These researchers compared modern (M) and historical (H) records of the Alpine Chipmunk (*Tamias alpinus)*. Gray spots are sampling locations where the species was not found. Black spots are sites where specimens were collected as part of the quantitative sampling effort. Red spots represent specimens collected or observed that were not part of the quantitative trapping effort.
   1. What can we interpret about the range of the chipmunk?

A screenshot of a cell phone

Description automatically generated

1. If historical records were incomplete or disorganized, what challenges would modern researchers encounter when attempting to track range shifts?
2. Think about your own lab notebook or field notebook. How might you change your habits to make sure the data you collect are usable in the future?

**References:**

GBIF.org, 2019. What is Darwin Core, and why does it matter? Available from: [https://www.gbif.org/darwin-core](https://www.gbif.org/darwin-core%20) [11 December 2019].

Moritz, C., J.L. Patton, C.J. Conroy, J.L. Parra, G.C. White, S.R. Beissinger. 2008. Impact of a Century of Climate Change on Small-Mammal Communities in Yosemite National Park, USA. Science 322(10): 261-264 and Supplementary Materials. DOI: 10.1126/science.1163428 S.

Museum of Vertebrate Zoology, University of California, Berkeley. 2015.<http://mvz.berkeley.edu/Grinnell.html>

USGS. 2019. Data Management. <https://www.usgs.gov/products/data-and-tools/data-management/data-standards>