



Challenges and trends in really **big** insect collection **data**sets

UCSB

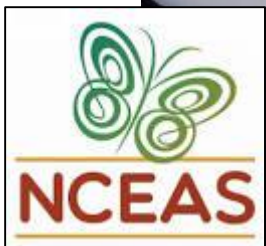
UNIVERSITY OF CALIFORNIA
SANTA BARBARA

Entomological Society of America
Big Data and Bugs Symposium
November 7, 2017

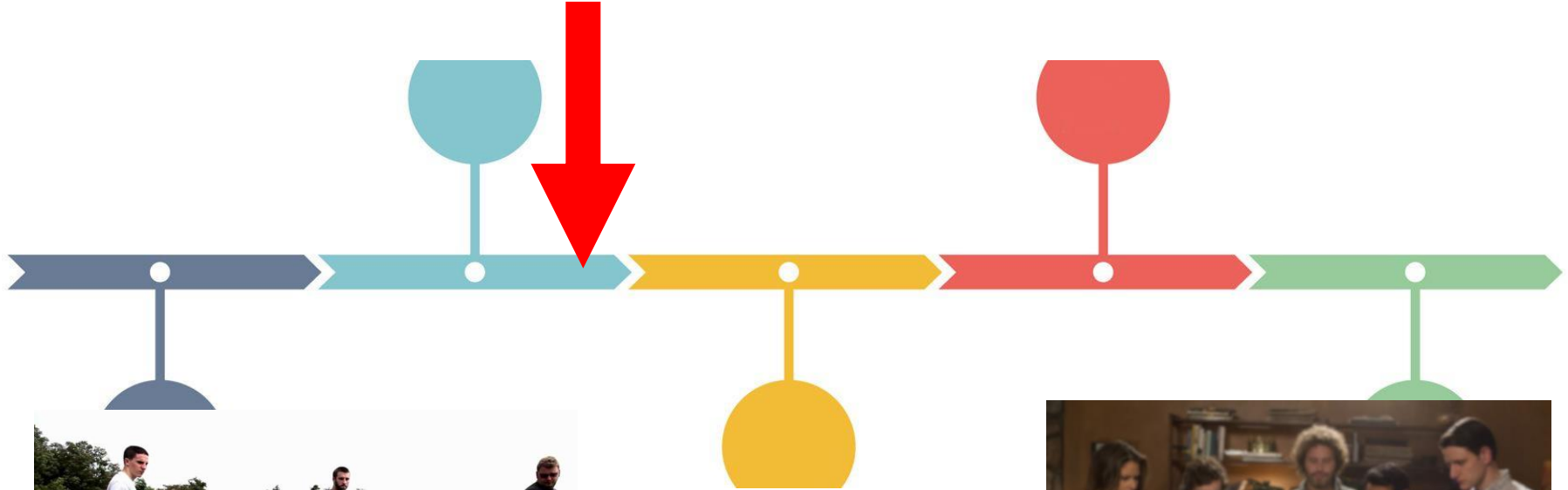
Katja C. Seltmann, PhD
Cheadle Center for Biodiversity and
Ecological Restoration
University of California Santa Barbara
seltmann@ccber.ucsb.edu



Georeferencing for Research
Use Workshop, Oct 4-7, 2016



You are Here





Home Search Images Fauna Projects Statistics Other Networks Work with SCAN Symbiota Contact Log In New Account Sitemap

Symbiota Collections of Arthropods Network (SCAN): A Data Portal Built to Visualize, Manipulate, and Export Species Occurrences

The Symbiota Collections of Arthropods Network (SCAN) serves specimen occurrence records and images from over 80 North American arthropod collections for **all** arthropod taxa. The focus is on North America but global in scope. SCAN is built on Symbiota, a web-based collections database system that is used for other taxonomic data portals, including (Symbiota Portals). SCAN is the primary repository for occurrence data produced by the three continuing Thematic Collections Networks (TCNs), the Southwest Collections of Arthropods Network (SCAN TCN), the Lepidoptera of North America Network (LepNet TCN), and arthropod data produced by InvertEBase TCN. InvertEBase serves occurrence data for mollusk and other non-arthropod taxa. We also host observational data, the largest data provider is iNaturalist. Each collection is primarily responsible for their data and we have structured the database to make it easy to include collections of interest when querying the database.

Important features of all Symbiota portals include:

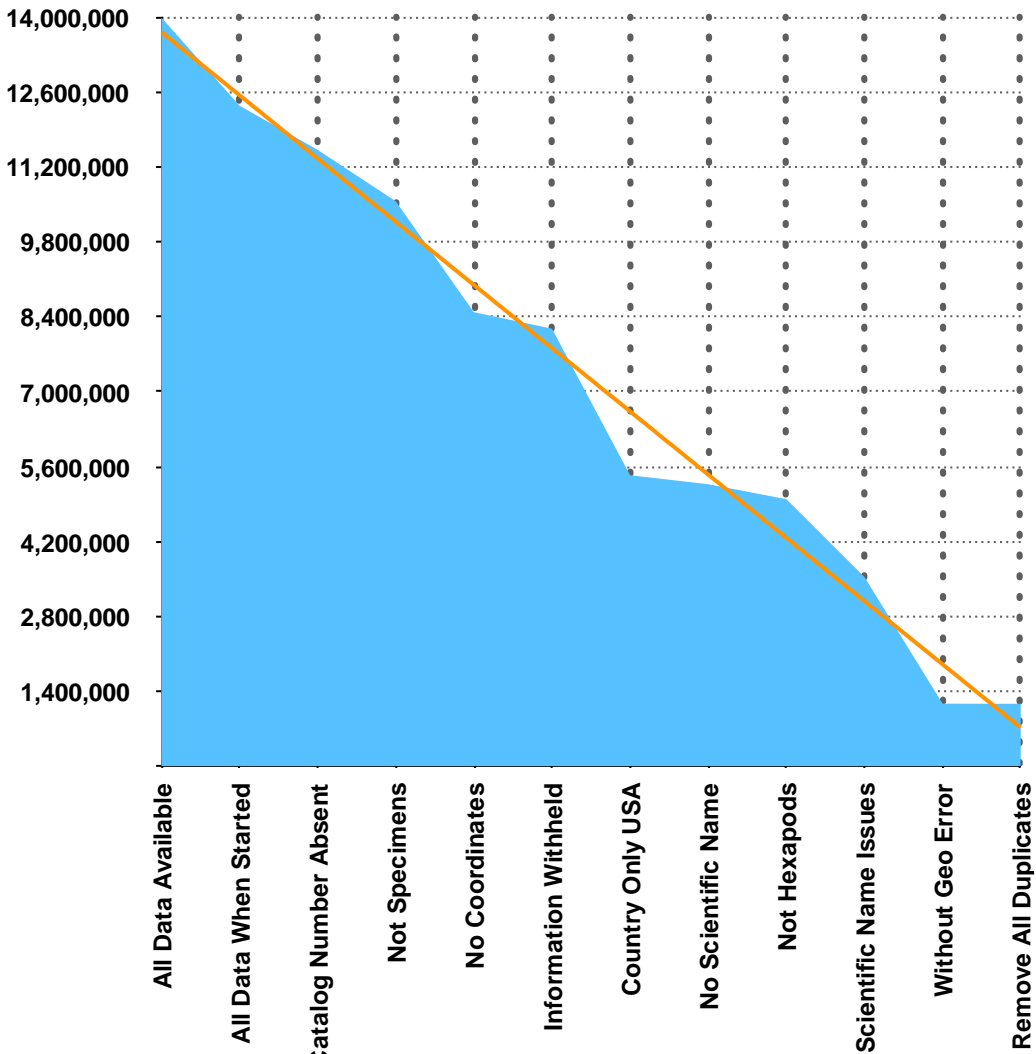
1. Easy web-based data entry.
2. Download entire datasets in two clicks.
3. Map georeferenced records in two clicks.
4. Upload high-resolution images & create species profile pages.
5. Design custom species lists for any locality at multiple scales.
6. Develop educational games with data.
7. Create taxonomic keys.

The key organizational feature is that each museum or project is listed as a separate collection, so that one database group does not interfere with another. End users can select all "collections", or just a subset. This website is the central data portal for SCAN; all other project information can be found on the LepNet WordPress site, including How-To-Guides and network updates.

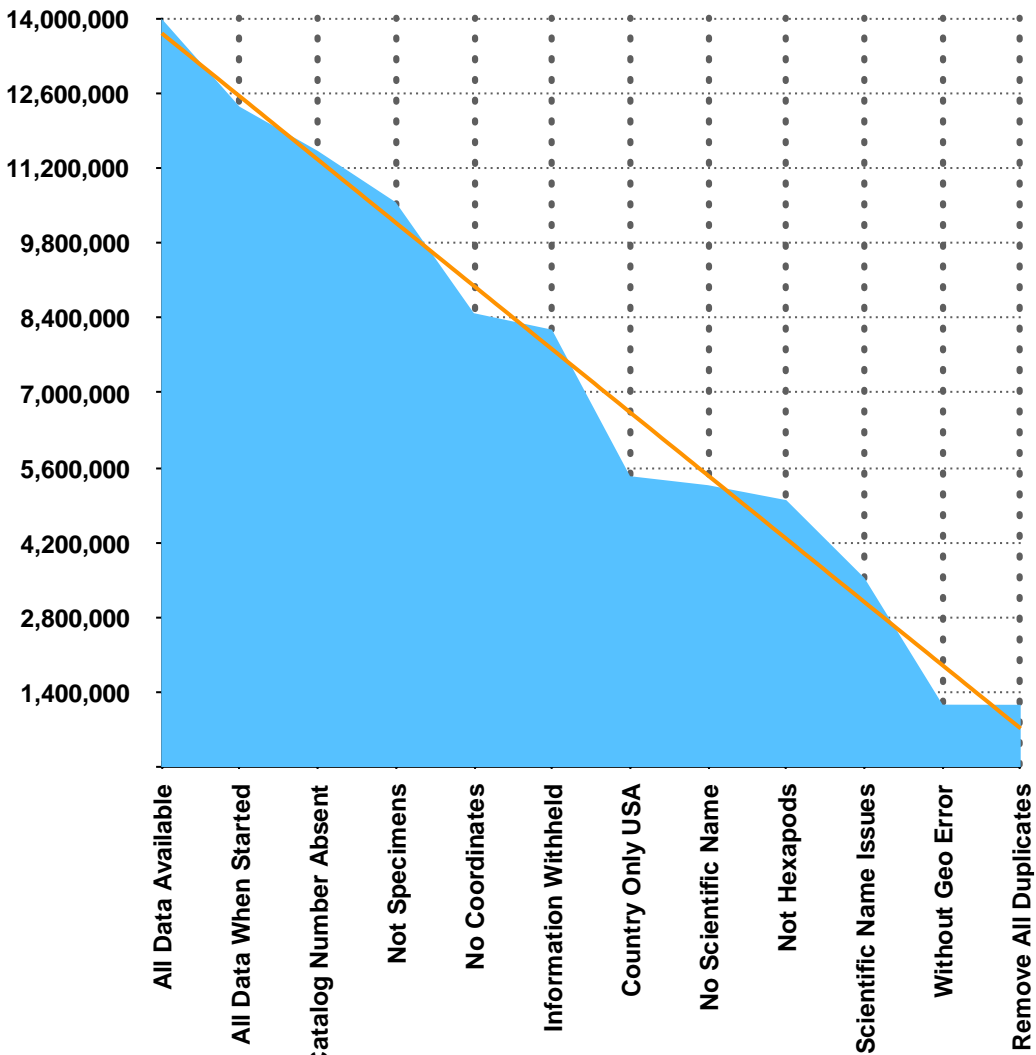
SCAN currently serves over 14 million records for 144,417 species, including 1,330,690 specimens imaged (8/1/17).

Taxon Search

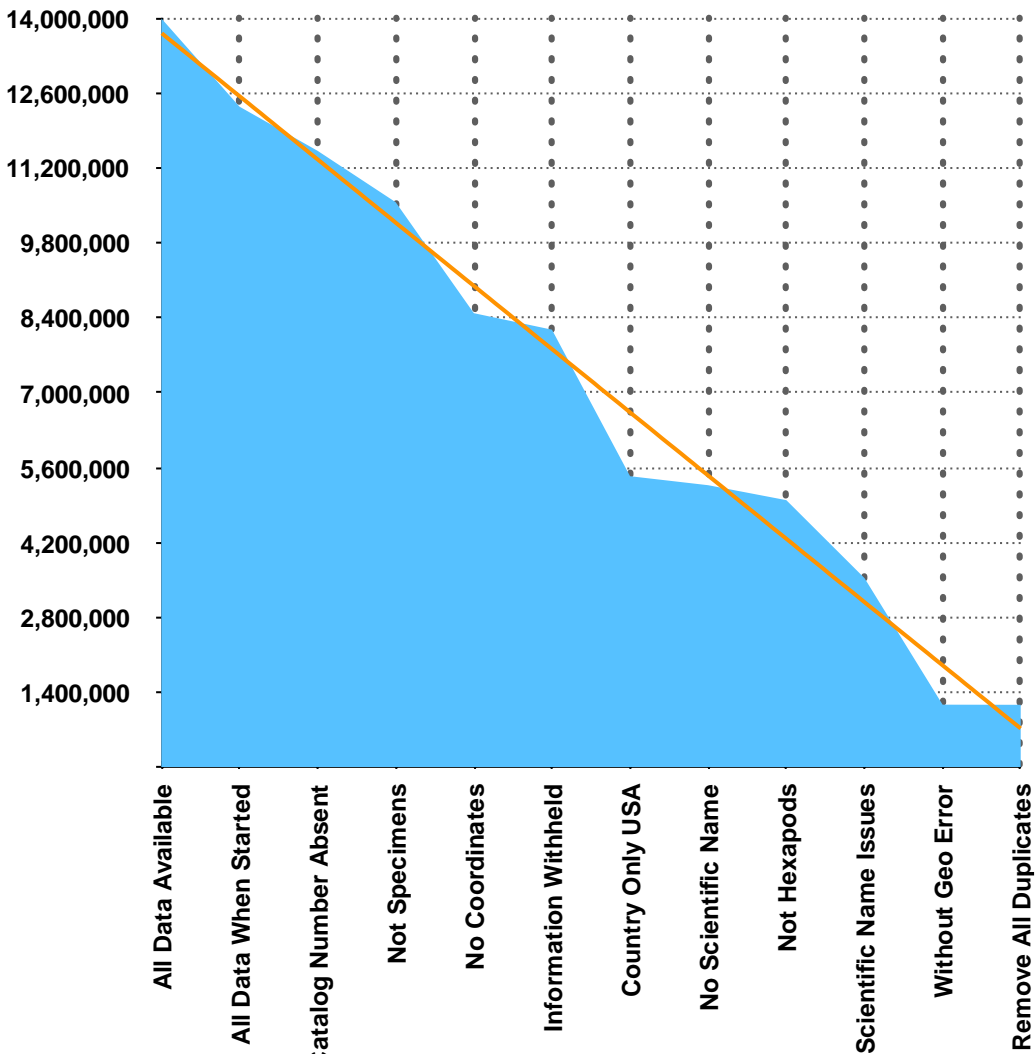
- 14,373,909 specimen records
- 79% georeferenced
- 10% imaged
- 58% identified to species
- 4,251 families
- 40,267 genera
- 221,890 species
- 228,206 total taxa (including subsp. and var.)



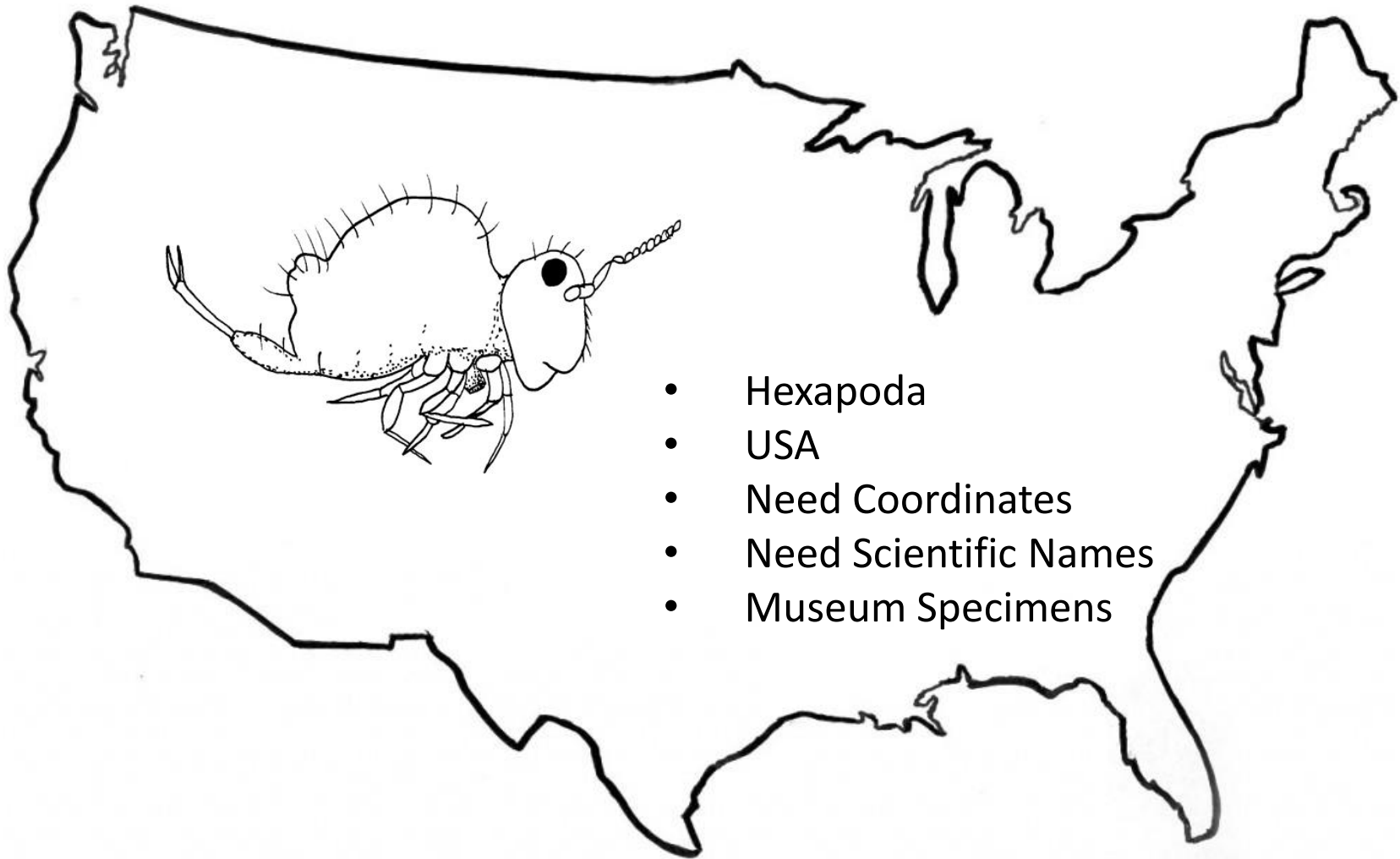
<i>Actions Taken</i>	<i>Data Available</i>
All Data Available	14,000,000
All Data When Started	12,347,785
Catalog Number Absent	11,523,828
Not Specimens	10,559,989
No Coordinates	8,485,230
Information Withheld	8,176,992
Not USA	5,435,585
No Scientific Name	5,268,055
Not Hexapods	4,989,736
Scientific Name Issues	3,532,798
Without Geo Error	1,162,978
Remove All Duplicates	1,161,064



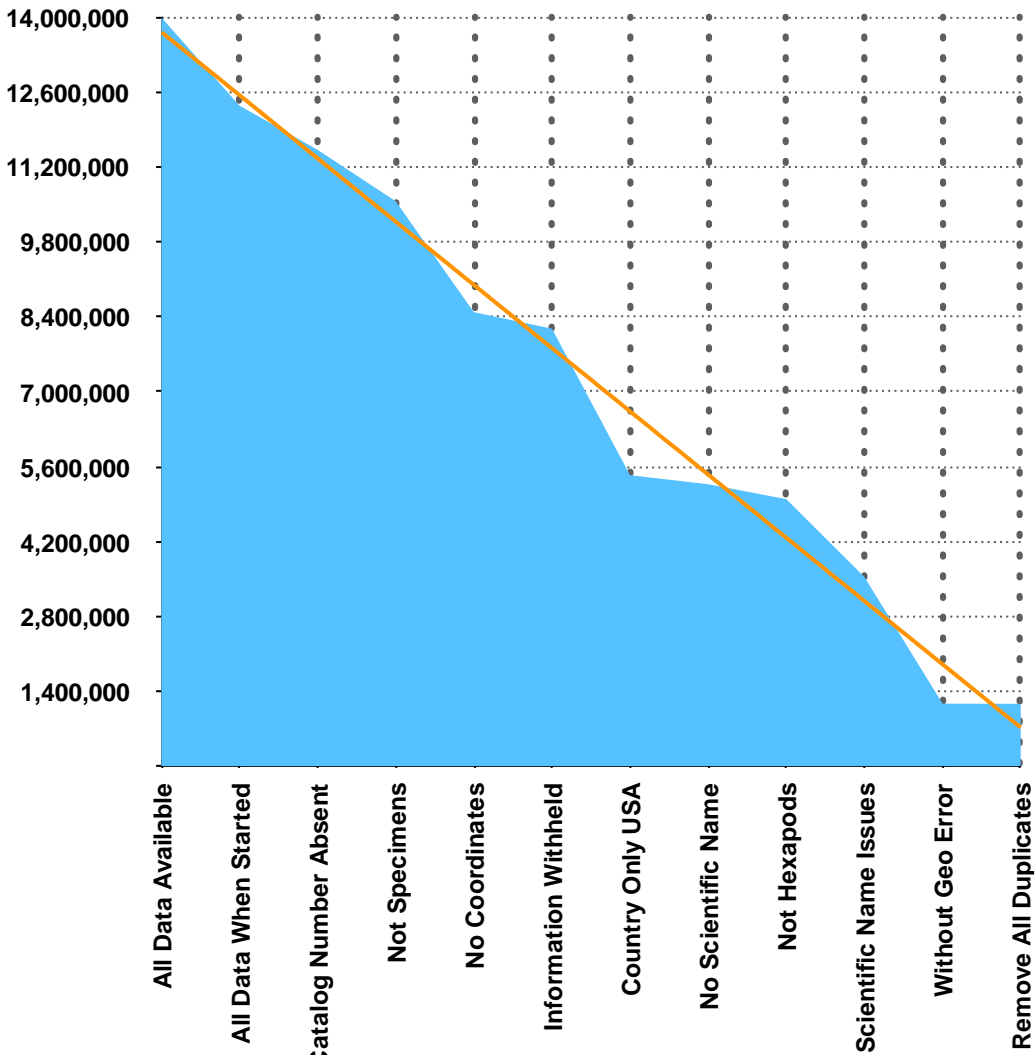
<i>Actions Taken</i>	<i>Data Available</i>
All Data Available	14,000,000
All Data When Started	12,347,785
Catalog Number Absent	11,523,828
Not Specimens	10,559,989
No Coordinates	8,485,230
Information Withheld	8,176,992
Not USA	5,435,585
No Scientific Name	5,268,055
Not Hexapods	4,989,736
Scientific Name Issues	3,532,798
Without Geo Error	1,162,978
Remove All Duplicates	1,161,064



<i>Actions Taken</i>	<i>Data Available</i>
All Data Available	14,000,000
All Data When Started	12,347,785
Catalog Number Absent	11,523,828
Not Specimens	10,559,989
No Coordinates	8,485,230
Information Withheld	8,176,992
Not USA	5,435,585
No Scientific Name	5,268,055
Not Hexapods	4,989,736
Scientific Name Issues	3,532,798
Without Geo Error	1,162,978
Remove All Duplicates	1,161,064



- Hexapoda
- USA
- Need Coordinates
- Need Scientific Names
- Museum Specimens

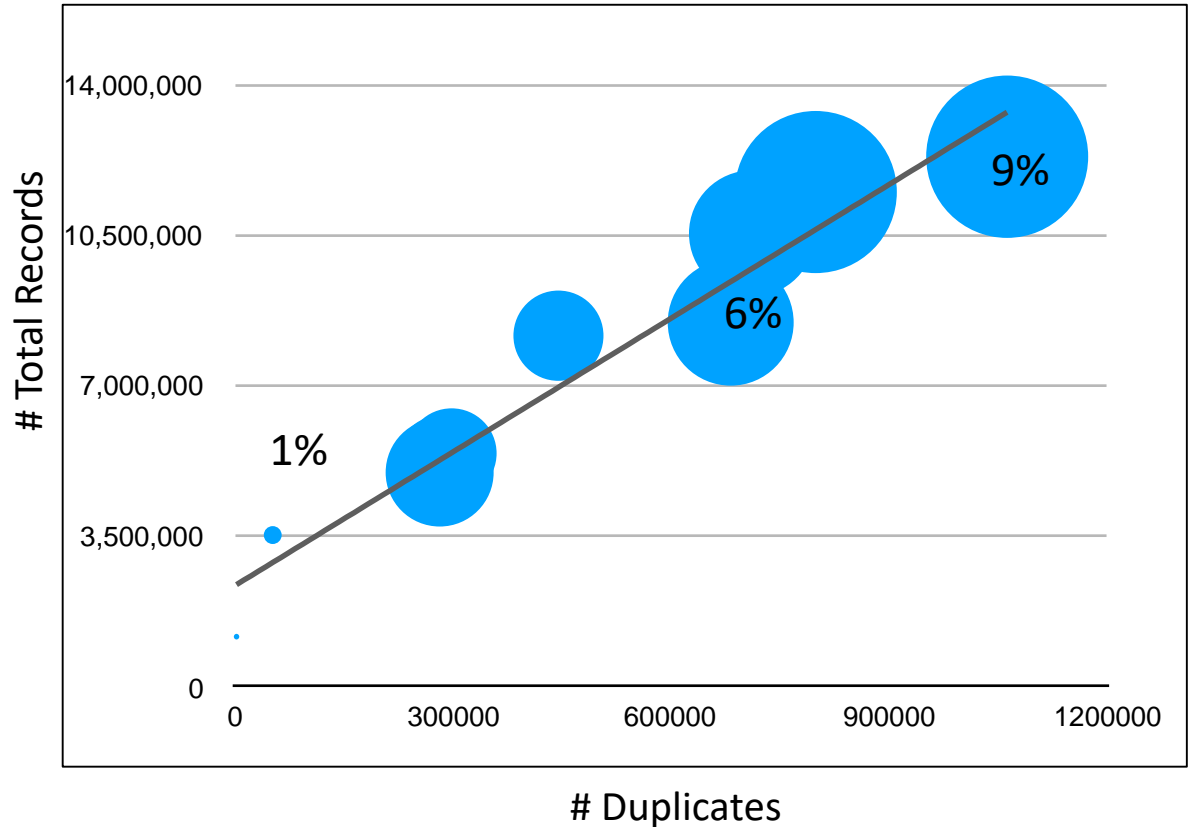


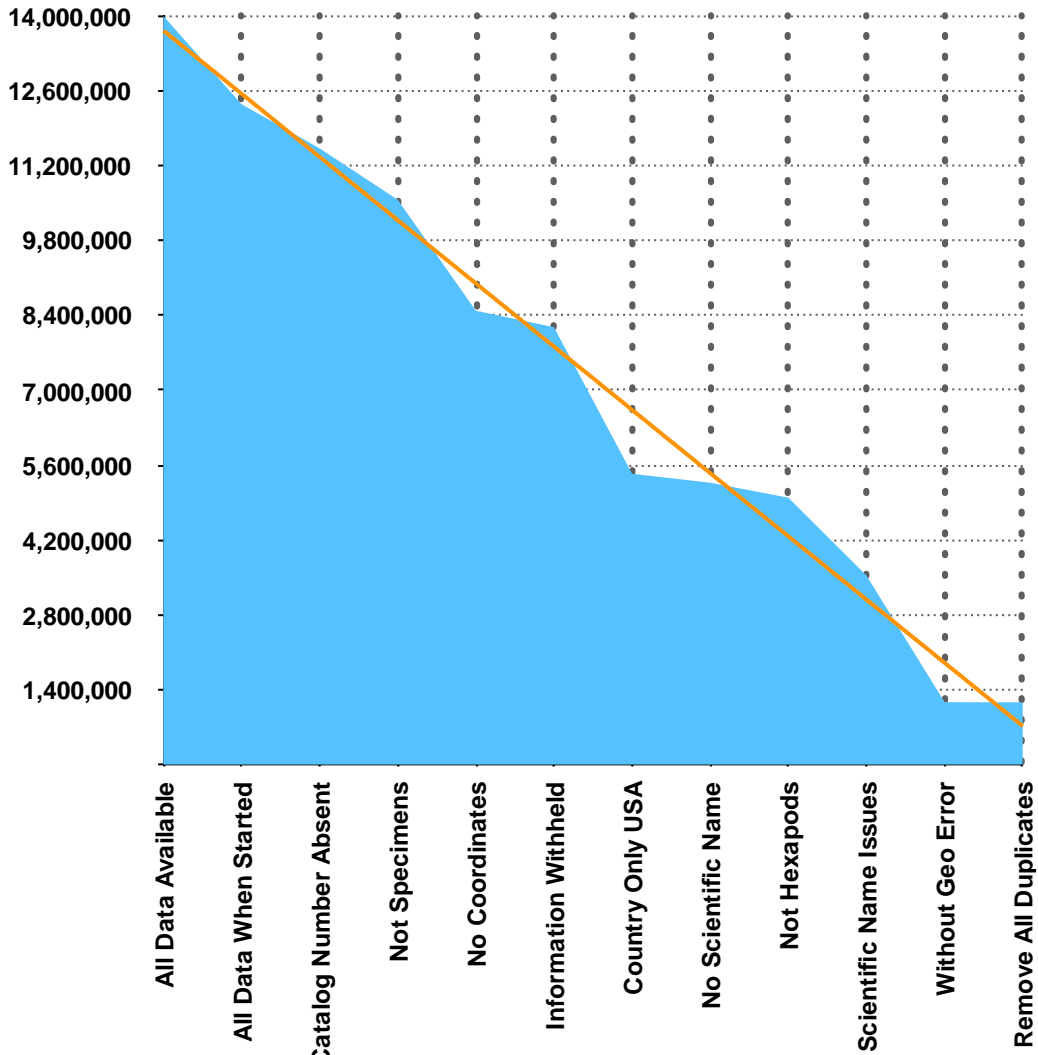

<i>Actions Taken</i>	<i>Data Available</i>
All Data Available	14,000,000
All Data When Started	12,347,785
Catalog Number Absent	11,523,828
Not Specimens	10,559,989
No Coordinates	8,485,230
Information Withheld	8,176,992
Not USA	5,435,585
No Scientific Name	5,268,055
Not Hexapods	4,989,736
Scientific Name Issues	3,532,798
Without Geo Error	1,162,978
Remove All Duplicates	1,161,064

Alpha Numeric Catalog Numbers



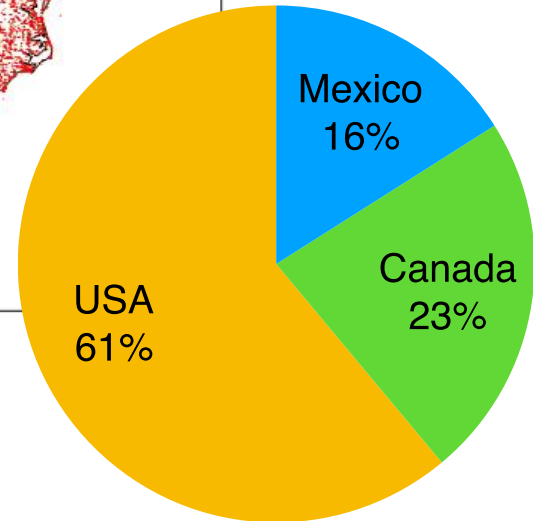
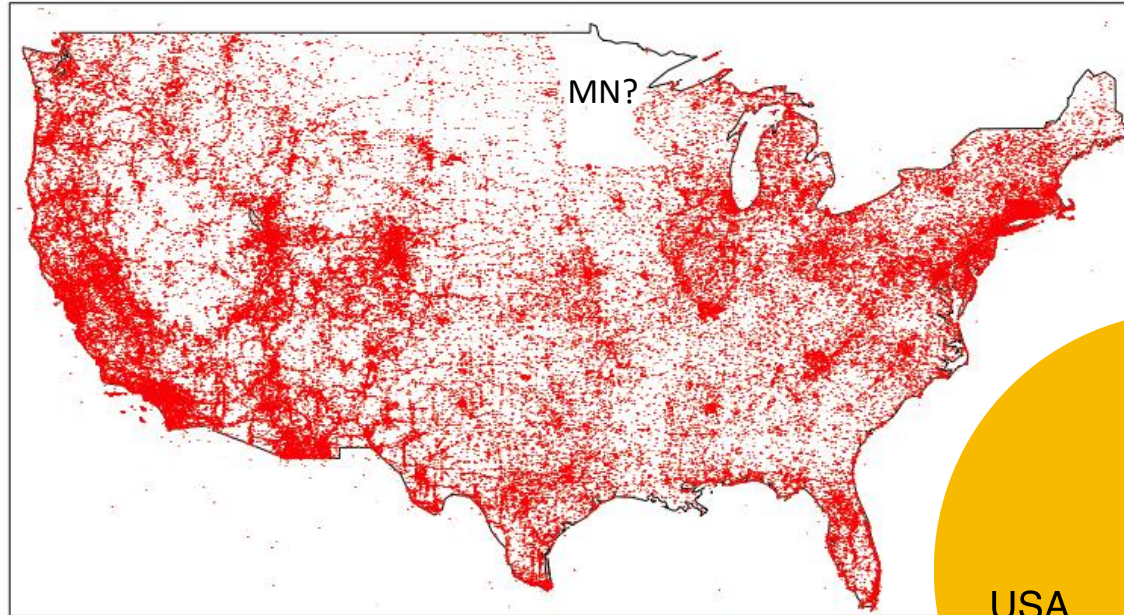
1,061,954 duplicates
=491 days to delete



<i>Actions Taken</i>	<i>Data Available</i>
All Data Available	14,000,000
All Data When Started	12,347,785
Catalog Number Absent	11,523,828
Not Specimens	10,559,989
No Coordinates	8,485,230
Information Withheld	8,176,992
Not USA	5,435,585
No Scientific Name	5,268,055
Not Hexapods	4,989,736
Scientific Name Issues	3,532,798
Without Geo Error	1,162,978
Remove All Duplicates	1,161,064

**No Coordinates
Information Withheld
Without Geo Error**



Term Name: coordinateUncertaintyInMeters

Identifier:	http://rs.tdwg.org/dwc/terms/coordinateUncertaintyInMeters
Class:	http://purl.org/dc/terms/Location
Definition:	The horizontal distance (in meters) from the given decimalLatitude and value empty if the uncertainty is unknown, cannot be estimated, or is n
Comment:	Examples: "30" (reasonable lower limit of a GPS reading under good conditions having 100 meter precision and a known spatial reference system). For
Details:	coordinateUncertaintyInMeters

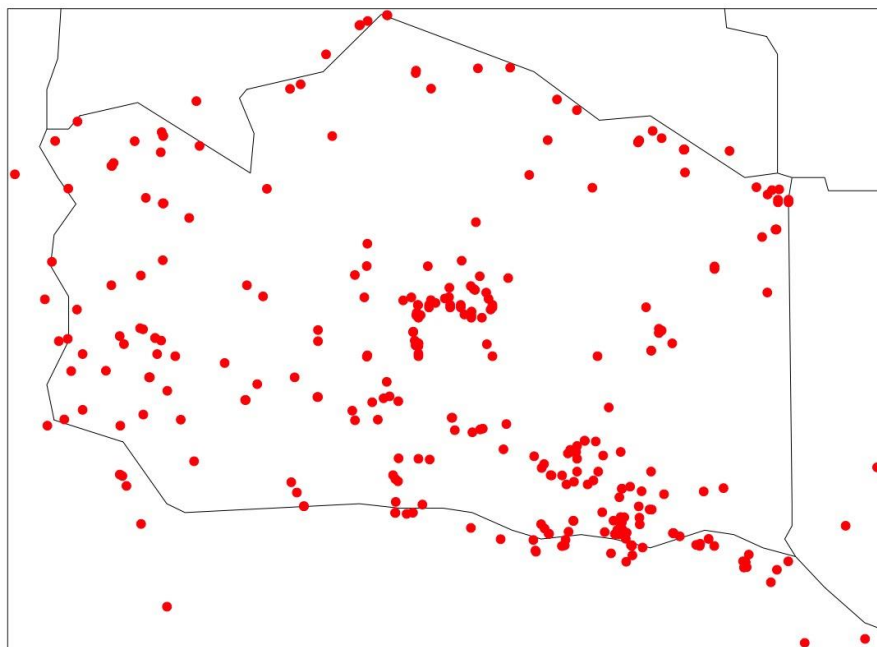
Term Name: coordinatePrecision

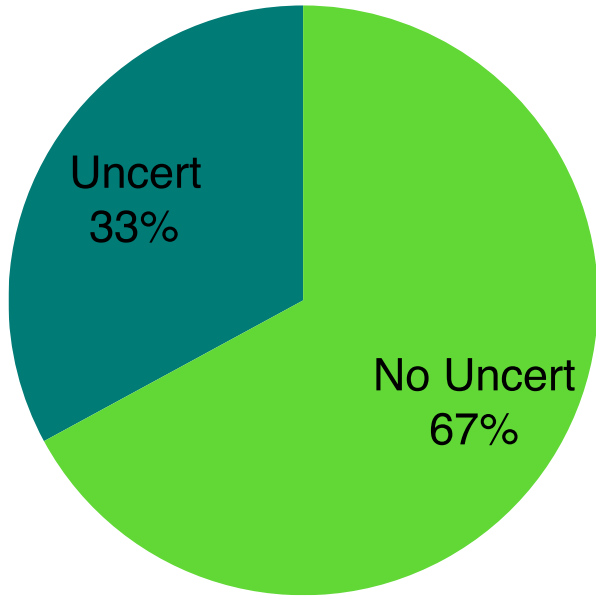
Identifier:	http://rs.tdwg.org/dwc/terms/coordinatePrecision
Class:	http://purl.org/dc/terms/Location
Definition:	A decimal representation of the precision of the coordinates given in the
Comment:	Examples: "0.00001" (normal GPS limit for decimal degrees), "0.00027"; http://terms.tdwg.org/wiki/dwc:coordinatePrecision
Details:	coordinatePrecision

Term Name: pointRadiusSpatialFit

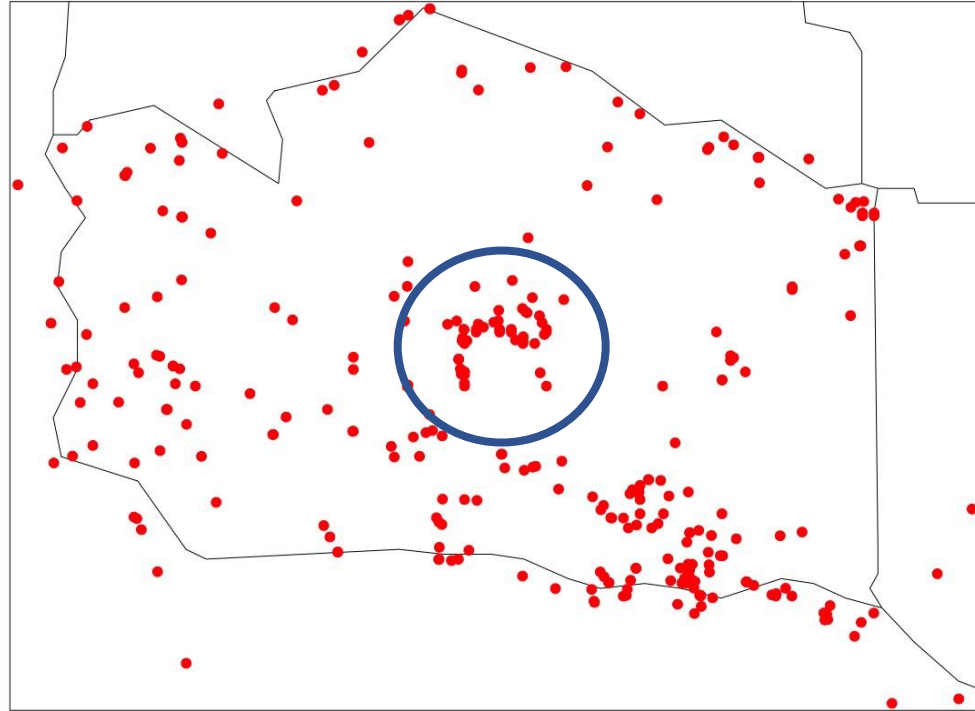
Identifier:	http://rs.tdwg.org/dwc/terms/pointRadiusSpatialFit
Class:	http://purl.org/dc/terms/Location
Definition:	The ratio of the area of the point-radius (decimalLatitude, decimalLongitude) representation of the Location. Legal values are 0, greater than or equal to the given point-radius does not completely contain the original representation is a point without uncertainty and the given georeference the same point, the pointRadiusSpatialFit is 1.
Comment:	Detailed explanations with graphical examples can be found in the "Guide"; http://terms.tdwg.org/wiki/dwc:pointRadiusSpatialFit
Details:	pointRadiusSpatialFit

**No Coordinates
Information Withheld
Without Geo Error**





**No Coordinates
Information Withheld
Without Geo Error**

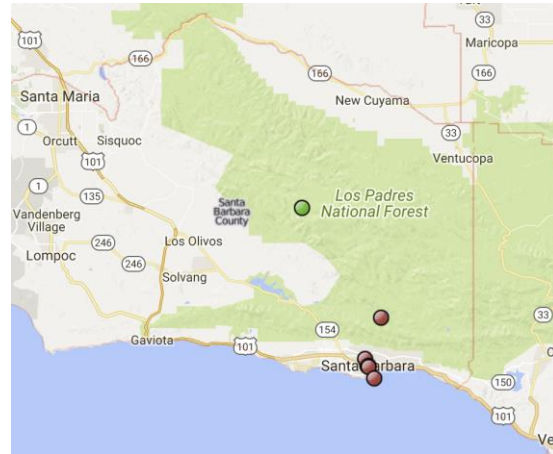


The problem with centroids without error

**No Coordinates
Information Withheld
Without Geo Error**



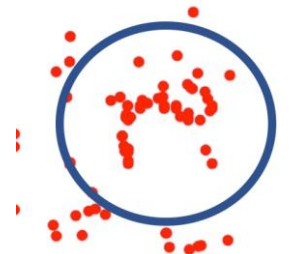
34.726867, -119.806743



34.733639, -119.856181

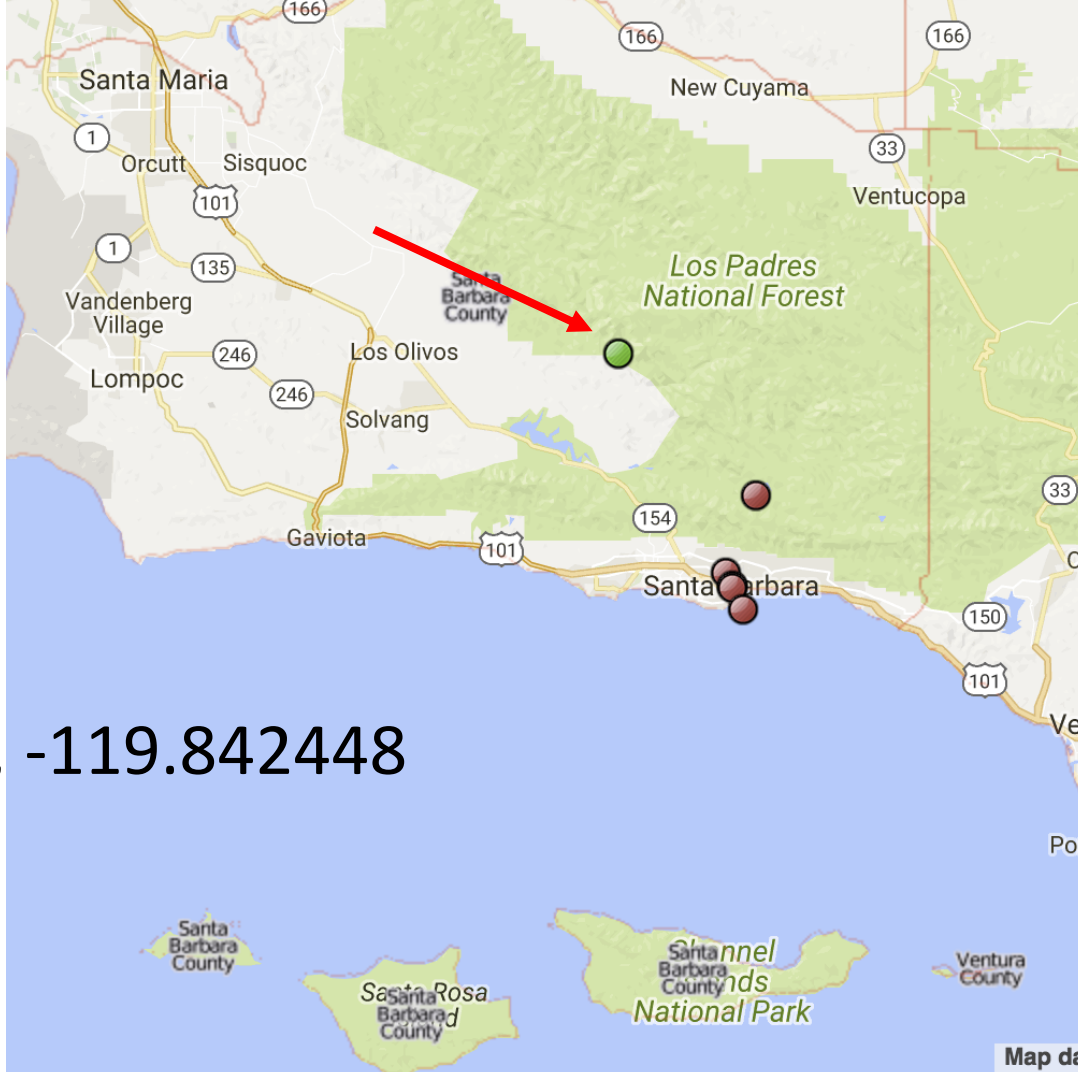


34.76072, -119.894633



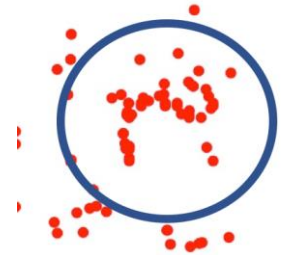
Centroid method of furthest point from the boundary of a county

Geographic center
(centroid)



34.665896, -119.842448

**No Coordinates
Information Withheld
Without Geo Error**



Map data

How do we get to Bigger Data?

At the Data Aggregator Level

- TDWG Data Quality Working Group
- Data quality across data providers

At the Research Level

- Understand present data cleaning considerations for large data.
 - Need new methods and repeatable practices
 - Calculate a hole in the middle that is likely the radius of a centroid and delete all records without an uncertainty

At the Data Provider Level

- Understand research practices
 - Provide extant for georeferences
 - Consider your data in the downstream context

When in
doubt,
we
throw it
out!

- Cheadle Center for Biodiversity
- NSF: Tri-Trophic TCN project PIs, digitizers, and collections
- Museum collections, herbaria, and curators worldwide
- Neil Cobb, Benjamin Brandt, Deb Paul, Jorrit Poelen, iDigBio and many others

Katja C. Seltmann, PhD

Cheadle Center for Biodiversity and Ecological Restoration

seltmann@ccber.ucsb.edu

