Challenges Using Biodiversity Data: examples of what to look for

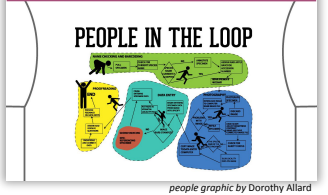
presenter:Arctic Data Center, DataONE, Environmental Data Initiative, ESIP, GBIF, iDigBio, NEON

#datahelpdesk
Ecological Society of America 2019 ESAUSSEE
Career Fair Center in the Exhibit Hall https://esa.org/louisville/career-fair/
Monday 12 August 230-300 PM
http://bit.ly/datahelpesa2019

Advancing the Digitization of Biological Collections

iDigBio Hub and Thematic (Museum) Collection Networks

(Meta)data Aggregator
Community Building
Collections Data



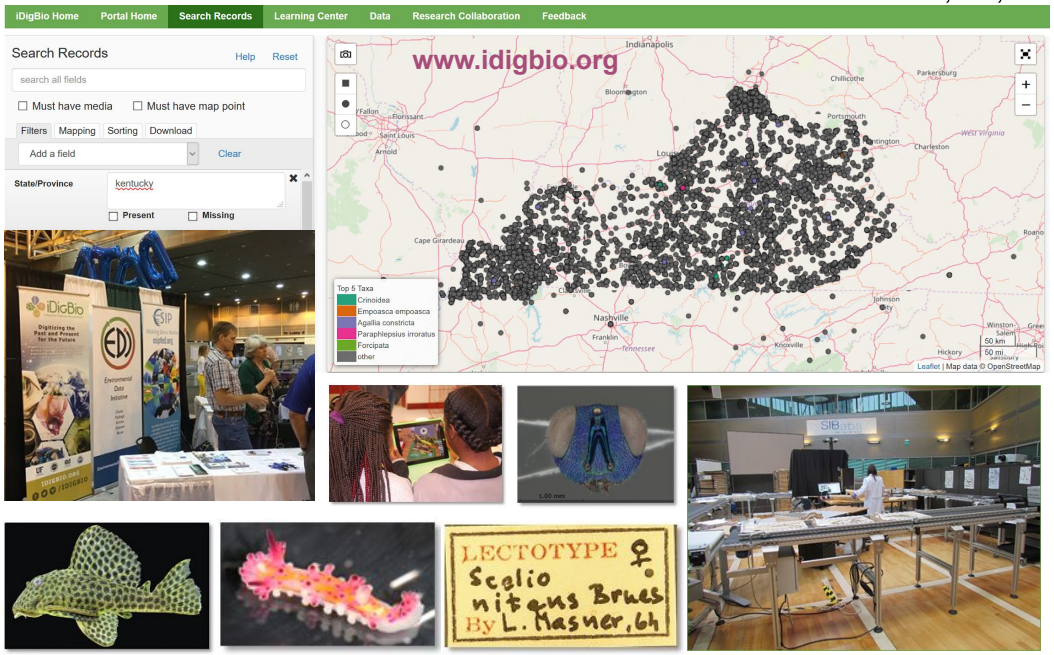
people graphic by Dorothy Allard

total: 119,768,942

Digitization
Workflows & Protocols
Dissemination

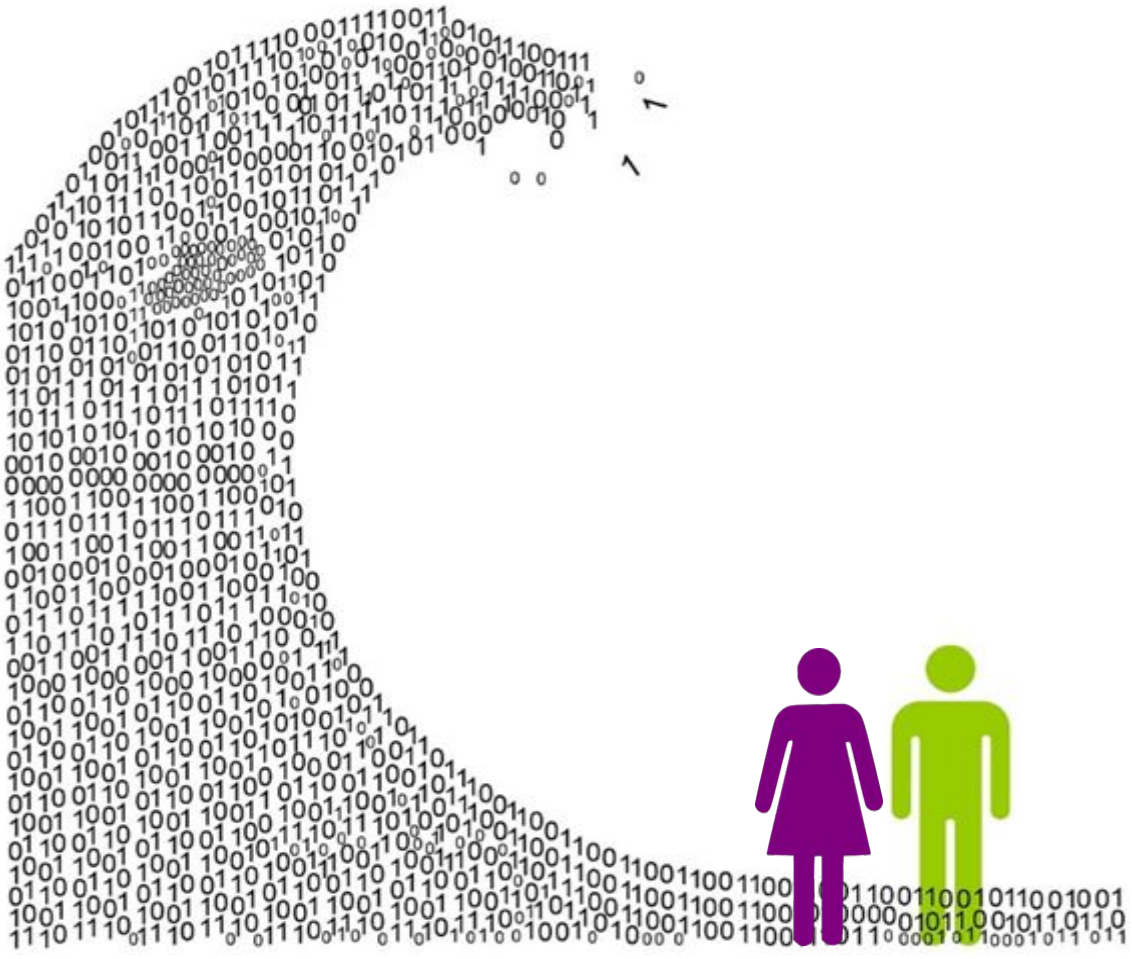
Research Use
Cyberinfrastructure
Tool collaboration
Portal development
ENM workshop
Research focus
Data quality
APIs

Training
Biodiversity informatics
Data skills and literacy
Collections software
Imaging
Project Management



Education Outreach
Citizen Science
K-12 materials
Undergraduate
Fossil Clubs
Mentor teachers

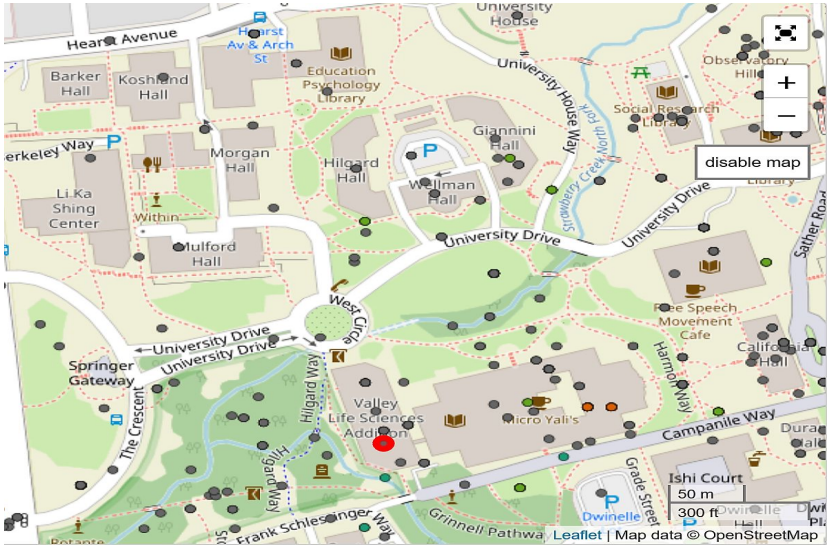
Methods
Workshops
Webinars
Symposia
Conferences
Working Groups
Short Courses
Adobe Connect
Listservs
Publications
Social Media @idigbio



Challenges researchers face

exercise: from the ecologist, collector, policy maker, or other downstream user of data (e.g. collections), what issues need to be addressed before applying your research methods?

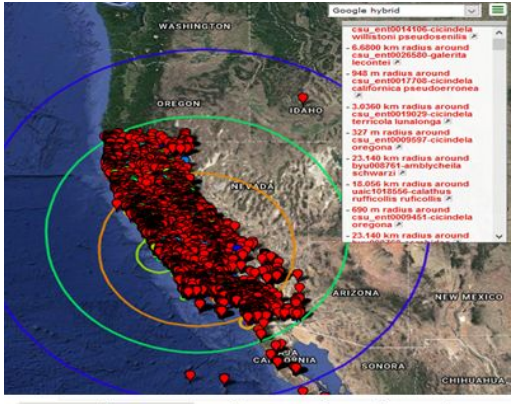
1. List three major challenges
2. Actions, tools, and influence



Data lessons compiled - inspired by workshop

Georeferencing for Research Use of Museum Collections Data

- Data mapped to standards
  - supports use and re-use (e.g. Darwin Core DwC, Ecological Metadata Language EML)
  - standards help with data validation and cleaning
- Data have issues
  - what are some you have experienced
  - need to be addressed before applying research methods
  - keep raw data raw
  - track your changes
- Issues can be grouped
  - Taxonomic / Nomenclatural
  - Locality / Place / Georeference
  - Time / Date
- Data visualization is key
  - QGIS lessons
  - Open Refine
  - R, etc.
- For future data - use data standards



id	idigbio:barcodeValue	dwc:basisOfRecord	dwc:bed	dwc:catalogNumber	dwc
515ab5a8-0c1a-4a53-9212-fd8a9070f2a9		preservedspecimen		osuc 474848	inse
52182933-ab95-442a-ba20-4d9333be21af		preservedspecimen		csu_ent0017464	inse
533c7312-3970-4e12-8190-917752c94899		preservedspecimen		csu ent0009279	inse
5515a8f4-0d6a-425e-a04a-172263ca3a84		preservedspecimen		csu_ent0014106	inse
559ee5d5-d60b-452d-873e-9f3133f62a85		preservedspecimen		osuc 479715	inse
5737e4dc-e600-4565-9676-93cd71913123		preservedspecimen		csu_ent0026580	inse
1c87cd60-c52d-4f05-b609-f8282a4f3869		preservedspecimen		osuc 471442	inse
1d0fd46b-1c6c-44bb-bd8c-96a12f22fe2a		preservedspecimen		asuhic0045568	inse
1f89aa2f-b34d-4b7c-8e5e-18f930882fc5		preservedspecimen		osuc 474825	inse
222770d9-2bb9-44e8-b477-13f45632f512		preservedspecimen		osuc 454183	inse
1393b0e2-82c3-42b1-bd74-90494e3c9a34		preservedspecimen		osuc 465205	inse
158bc3d9-0c37-4c90-8c2c-2c8b948c8e28		preservedspecimen		csu_ent0017708	inse
15c78b96-f4bb-4e4a-8972-4922f82eabc3		preservedspecimen		osuc 475101	inse

Carabidae (beetles) of California

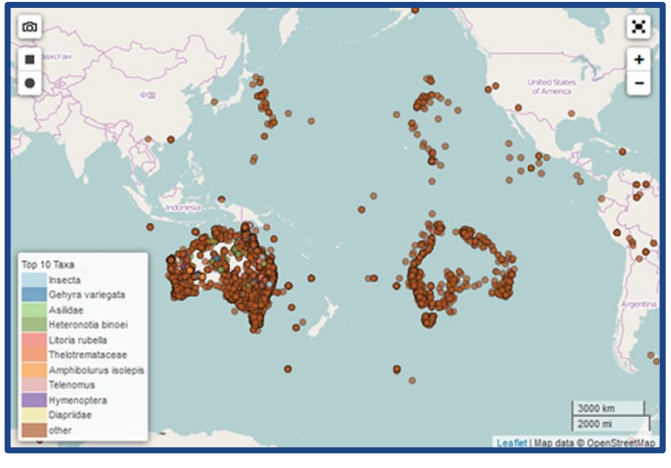
Synthesis of issues

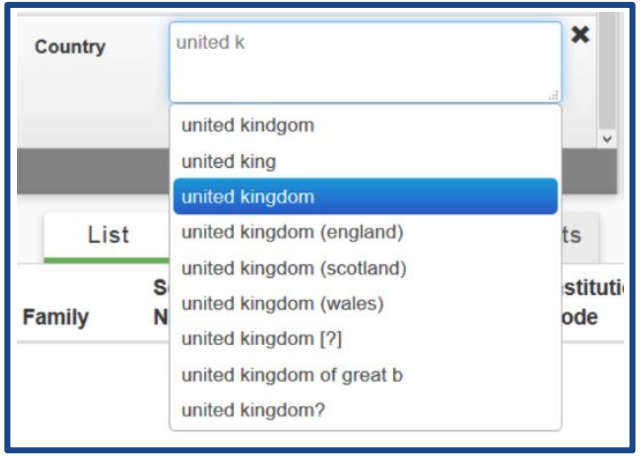
evaluating the research fitness-for-use of these data
creating a list of data quality checks

- Timey-wimey stuff
  - date issues like formats
- Geography
  - place name issues
  - out of expected bounds
  - missing metadata
- Taxonomy
  - taxon name issues*
  - concepts
  - authority files
  - parsing

Grouping data issues

- time (date)
- locality
- place





Date and time issues

Following

From a @HydraInABox interview: "People will put anything and their dog in the date field. It's absolutely astonishing."

The date field (and others)

1. Order matters for dates
   - use yyyy-mm-dd
2. What to do when you have no value to share?
   - leave it blank
   - avoid placeholder values
3. 0 has meaning

dwc:eventDate	dwc:eventDate
1900-01-01	1900-01-01
1900-04-01	1900-04-01
03-04-03	1903-04-03
1901-08-17	1901-08-17
1901-08-17	1901-08-17
0	
999	
1903-05-02	1903-05-02
1903-05-02	1903-05-02
1903-05-02	1903-05-02

Season of observation / collection

1. Does the date fit the organism?
2. Is it an outlier?
3. Or an error?

CM	CN
dwc:eventDate	dwc:scientificName
1900-01-01	calosoma semilaeve
1900-04-01	platynus brunneomarginatus
1903-04-03	calathus ruficollis ruficollis
1901-08-17	cicindela trifasciata sigmoidea
1901-08-17	cicindela trifasciata sigmoidea
	cicindela senilis
	cicindela senilis
1903-05-02	omus tularensis
1903-05-02	omus californicus
1903-05-02	omus tularensis

From gene names to dates?

Ziemann *et al. Genome Biology* (2016) 17:177
DOI 10.1186/s13059-016-1044-7

Genome Biology

COMMENT Open Access

Gene name errors are widespread in the scientific literature

Mark Ziemann[1], Yotam Eren[1,2] and Assam El-Osta[1,3*]

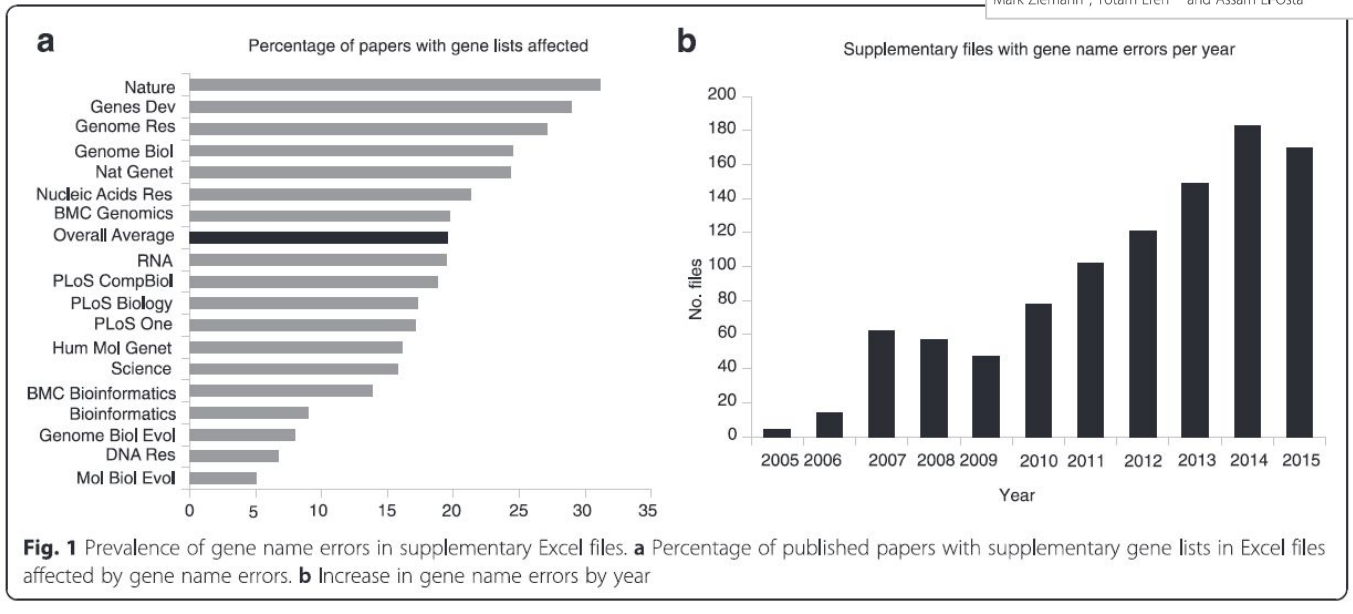
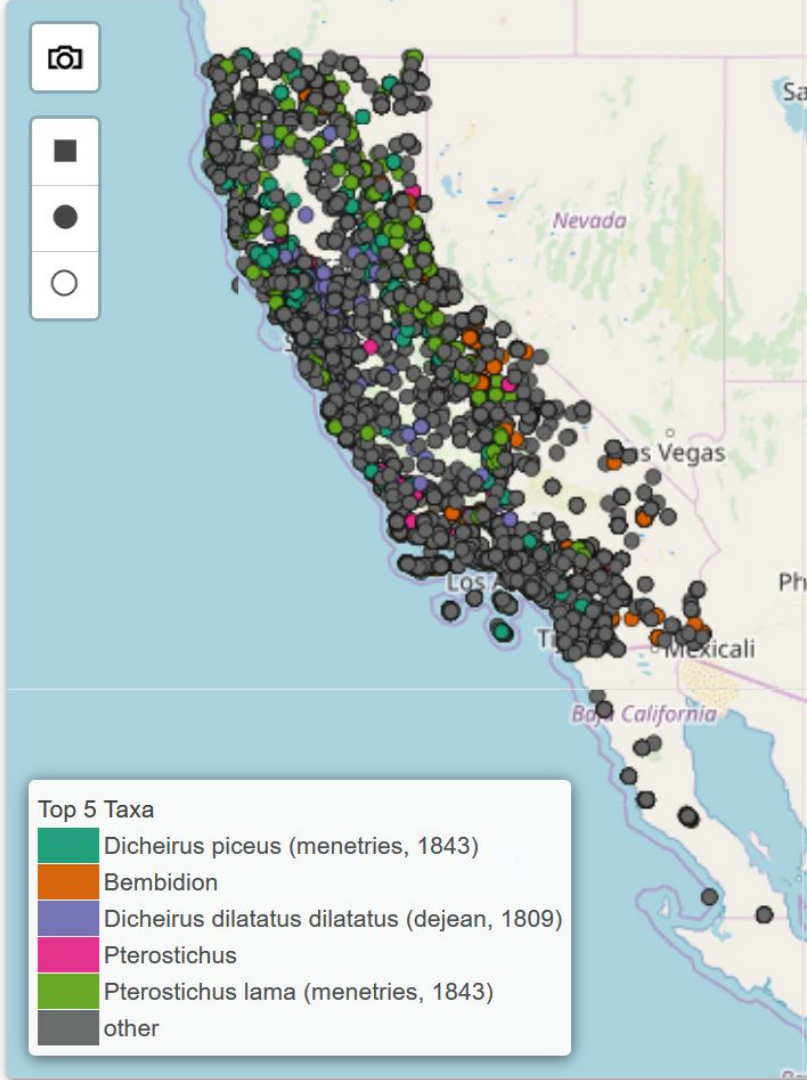


Fig. 1 Prevalence of gene name errors in supplementary Excel files. **a** Percentage of published papers with supplementary gene lists in Excel files affected by gene name errors. **b** Increase in gene name errors by year

Taxonomic names

1. Check spatial bounds for taxa
    a. expected or not?
2. Check downloaded data
    a. raw and tweaked data comparison needed



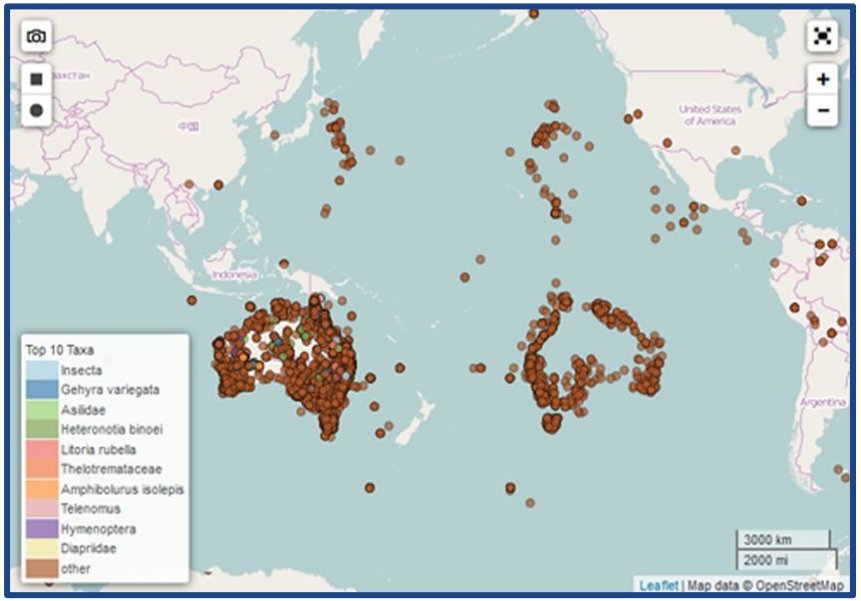
What level of taxonomic determination is required for your research question?

1. Remove records with undetermined, indet, empty
2. Use **dwc:taxonRank** to remove data outside ranks needed
3. Check raw and automated taxonomic names
4. Check endings - are they really the same?
5. Check synonymies, decide which names to use

CN	CO
dwc:scientificName	dwc:taxonRank
calosoma semilaeve	species
platynus brunneomarginatus	species
calathus ruficollis ruficollis	subspecies
carabidae	family
cicindela trifasciata sigmoidea	subspecies
cicindela senilis	species
indet.	
omus tularensis	species
omus californicus	species
omus	genus
omus tularensis	species
omus tularensis	species
pterostichus lama (menetries, 1843)	species

Got geopoints?

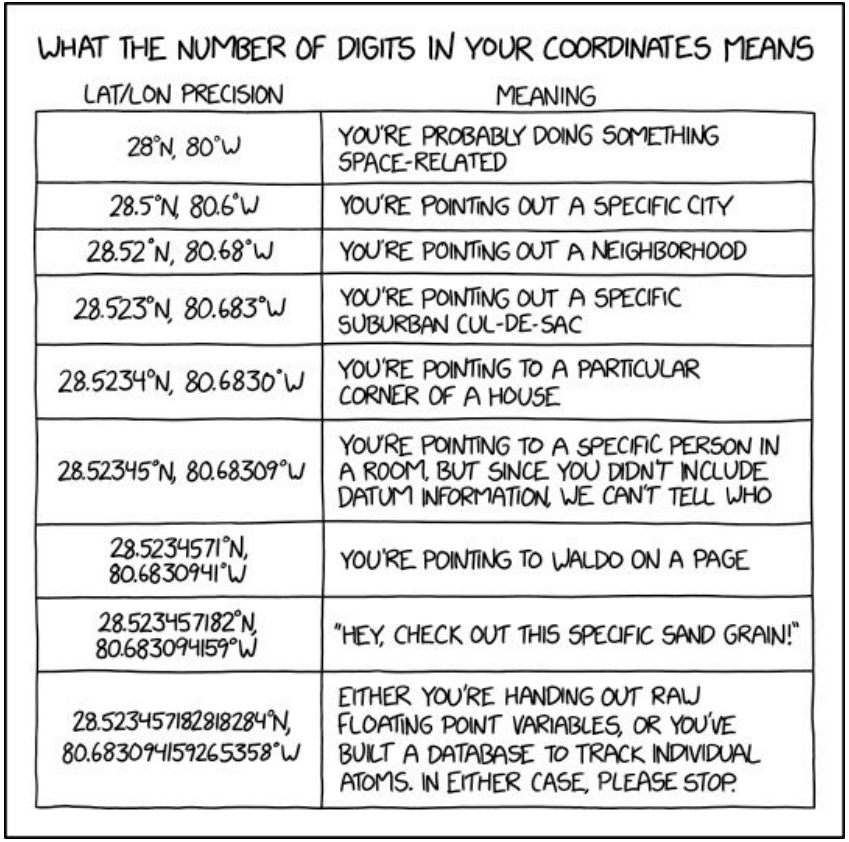
1. Map them
2. use layers to check reasonableness
   a. bounding boxes
   b. habitat, terrain
3. look for outliers and transpositions
4. Place could suggest cultivated, port, or zoo record



Presence data or precise data?

1. How fine-grained of a georeference do you need for your research question/model?

2. How many decimal places should make you suspect a data point? (fake precision)



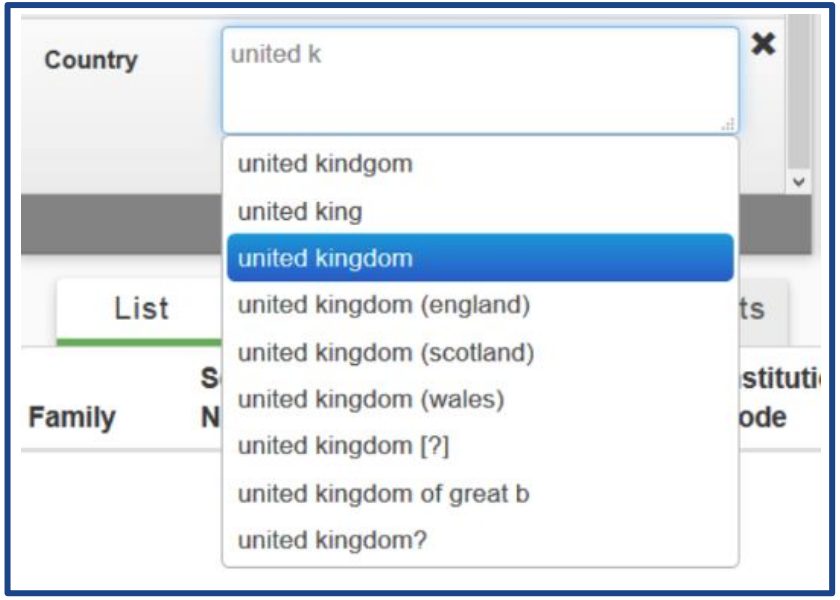
WHAT THE NUMBER OF DIGITS IN YOUR COORDINATES MEANS

LAT/LON PRECISION	MEANING
28°N, 80°W	YOU'RE PROBABLY DOING SOMETHING SPACE-RELATED
28.5°N, 80.6°W	YOU'RE POINTING OUT A SPECIFIC CITY
28.52°N, 80.68°W	YOU'RE POINTING OUT A NEIGHBORHOOD
28.523°N, 80.683°W	YOU'RE POINTING OUT A SPECIFIC SUBURBAN CUL-DE-SAC
28.5234°N, 80.6830°W	YOU'RE POINTING TO A PARTICULAR CORNER OF A HOUSE
28.52345°N, 80.68309°W	YOU'RE POINTING TO A SPECIFIC PERSON IN A ROOM, BUT SINCE YOU DIDN'T INCLUDE DATUM INFORMATION, WE CAN'T TELL WHO
28.5234571°N, 80.6830941°W	YOU'RE POINTING TO WALDO ON A PAGE
28.523457182°N, 80.683094159°W	"HEY, CHECK OUT THIS SPECIFIC SAND GRAIN!"
28.5234571828182841°N, 80.6830941592653581°W	EITHER YOU'RE HANDING OUT RAW FLOATING POINT VARIABLES, OR YOU'VE BUILT A DATABASE TO TRACK INDIVIDUAL ATOMS. IN EITHER CASE, PLEASE STOP.

Coordinate Precision from xkcd https://xkcd.com/2170/

Names for people, and in this case places

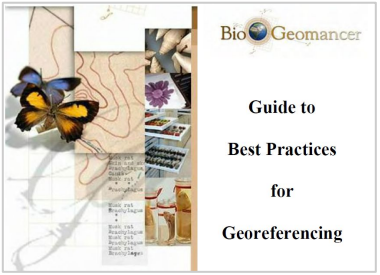
- Country names as example
  - check multiple values
  - check for abbreviations, etc.



Time travel to love your past self

use data standards

- Darwin Core (dwc)
- Ecological Metadata Language (EML)
- Best practices guide to georeferencing
- QGIS lesson for visualizing data to before analysis
- ...ask about more tools and resources at the #datahelpdesk in the exhibit hall



Collections Biodiversity Data - *expected and emerging uses*

Important Human Uses

- Evolutionary medicine,
- Disease discovery, tracking, and treatment
- Food security,
- Biodiversity conservation and sustainability,
- Computation,
- Design,
- Evolution and justice,
- Development of new types of biodiversity theories that accommodate newly emerging data streams.

Nelson G, Ellis S. 2018. The history and impact of digitization and digital data mobilization on biodiversity research. Philosophical Transactions of the Royal Society B Volume 374 Issue 1763 https://doi.org/10.1098/rstb.2017.0391

Emerging Research Angles

- Supplementing existing datasets with digital layers to enhance niche and species distribution modeling,
- Use of 3D/CT data for generating and testing new hypotheses,
- Implementation of convolutional neural networks (CNN) and deep learning in the analysis of image,
- Data for taxonomic determination and specimen curation,
- Delineation of traits in specimen images,
- Determination and identification to genus or species from, sediment-deposited pollen grains.

ESAUSSEE Data Help Desk

Who we are and how to find us (in the exhibit hall)

Amber Budden, @aebudden, @DataONEorg, aebudden@nceas.ucsb.edu
Christina Alba, @botanic, christina.alba@botanicgardens.org
Christine Laney, @cmlaney, @NEON_sci, claney@battelleecology.org
Deborah Paul, @idbdeb, @idigbio, dpaul@fsu.edu
Dmitry Schigel, @dschigel, @GBIF, dschigel@gbif.org
Jeanette Clark, @sjeanetteclark, @ArticDataCtr, jclark@nceas.ucsb.edu
Karl Benedict, @kbene, @ESIPfed, kbene@unm.edu, president@esipfed.org
Kelsey Yule, @kmyule, biorepo.neonscience.org, kmyule@asu.edu
Kristin Vanderbilt, @vanderbik, @EDIgotdata, krvander@fiu.edu
Kyle Copas, @kylecopas, @GBIF, kcopas@gbif.org
Laura Brenskelle, @lbrensk, @idigbio, lbrensk@ufl.edu
Margaret O'Brien @obrienmobb, @EDIgotdata, margaret.obrien@ucsb.edu
Megan Jones, @MeganAHJones, @NEON_sci, mjones01@battelleecology.org
Nico Franz, @taxonbytes, biorepo.neonscience.org, nico.franz@asu.edu
Rebekah Wallace, www.eddmaps.org, bekahwal@uga.edu
Stevan Earl, @StevanEarl, stevan.earl@asu.edu
William Michener, @DataONEorg, william.michener@gmail.com