

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>



NSF
ARCTIC
Data
Center

DataONE



ESIP
MAKING DATA MATTER



GBIF



neon
Operated by Battelle



(Meta)data Aggregator

Community Building

Collections Data

PEOPLE IN THE LOOP

people graphic by Dorothy Allard

Advancing the Digitization of Biological Collections

iDigBio Hub and Thematic (Museum) Collection Networks

total: 119,768,942

Digitization

Workflows & Protocols
Dissemination

Research Use

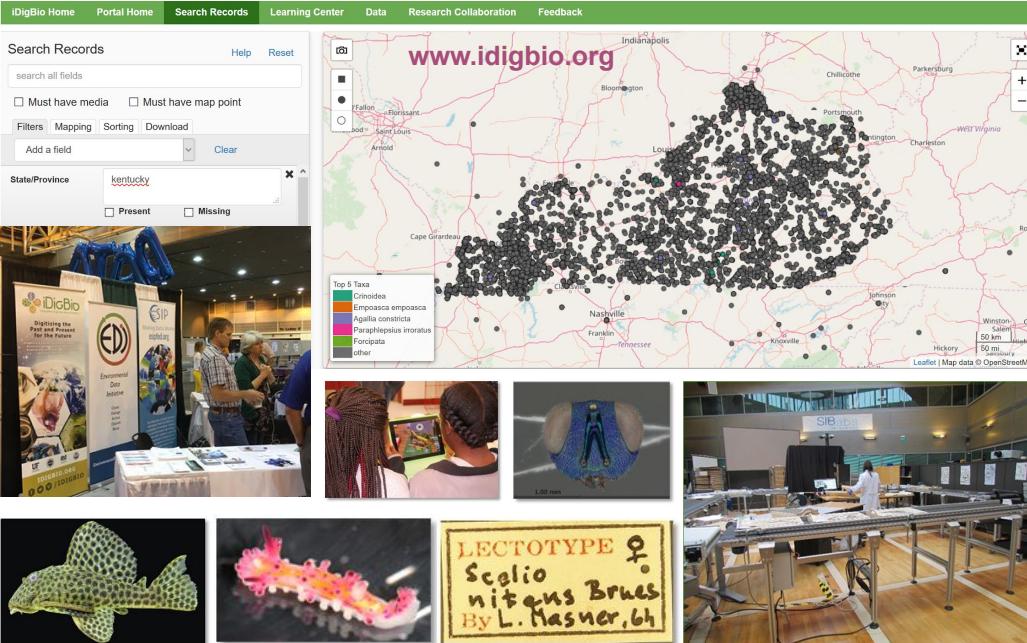
Cyberinfrastructure
Tool collaboration
Portal development
ENM workshop

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

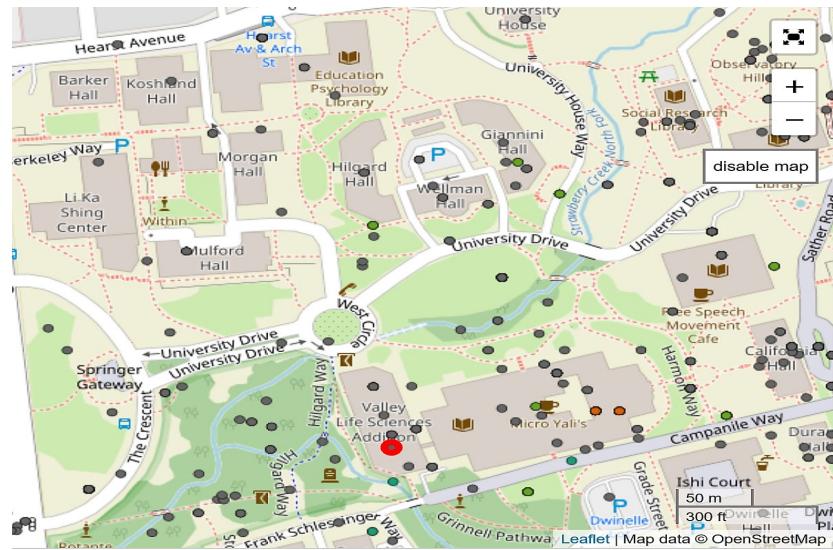
A large binary matrix (0s and 1s) representing gender data. The matrix is oriented diagonally, with the top-left corner at the top and the bottom-right corner at the bottom. A purple female icon and a green male icon are overlaid on the matrix, positioned in the lower-right quadrant. The matrix contains numerous binary digits, with a significant portion of the data being zeros.

Gender	Count
Female	~1500
Male	~1500

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



A	B	C	D	E
idigbio:barcodeValue	digicat:Record	dwc:bd	dwc:catalogNumber	dwc:isPresent
2. 5156b5d8-01ca-4a53-9211-f68e07079e69	preservespecimen	cu_entr00274848	inse	
3. 5212933-3a9f-442a-ba30-4933ba21af	preservespecimen	cu_entr0017464	inse	
4. 533c7312-3970-4e12-b150-917752d48499	preservespecimen	cu_entr0029729	inse	
5. 5515a8f4-0d94-425e-a04a-172263ca3b84	preservespecimen	cu_entr0014106	inse	
6. 559ee5d5-4600-452d-873e-9f113fe2a85	preservespecimen	osuc_479715	inse	
7. 57374edc-e600-4565-9676-91cd71913123	preservespecimen	cu_entr0026580	inse	
8. 1c87fd60-5240-4f05-9609-f8282a4f3886	preservespecimen	osuc_471442	inse	
9. 1d05446b-11c-44bb-bd8c-96a1272f7e2a	preservespecimen	asubch00045568	inse	
10. 1d05446b-11c-44bb-bd8c-96a1272f7e2a	preservespecimen	osuc_454183	inse	
11. 22277009-2b0a-44e8-ba77-33415632512	preservespecimen	osuc_545183	inse	
12. 1393b0e2-43c3-42b1-bd74-90494a3c934	preservespecimen	osuc_465205	inse	
13. 158bcbdf-0c37-4c90-8c2c-2c0fb948de28	preservespecimen	cu_entr00217708	inse	
14. 15c78b96-14bb-4e4a-8972-4922982zeabc3	preservespecimen	osuc_475301	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

doi: 10.3897/rio.4.e32449

GRU Workshop Conversation on Data Quality Considerations and Checks.

Data quality (dq) considerations and checks – an annotated bibliography.

Workshop participants discussed the following topics:

RIO Home Articles About About Pensoft Books E-Books Blog journals Register | Log in

Workshop Report Research Ideas and Outcomes 4: e32449 doi: https://doi.org/10.3897/rio.4.e32449 (17 Dec 2018) Reviewed v1 Unreviewed XML PDF Twitter Facebook LinkedIn Email Contents Article info Citation Metrics Reviews Related Figs Tables Data Refs Cited

Georeferencing for Research Use (GRU): An integrated geospatial training paradigm for biocollections researchers and data providers

Research Ideas and Outcomes 4: e32449 doi: 10.3897/rio.4.e32449 Received: 14 Dec 2018 | Published: 17 Dec 2018

Authors

Katja C. Seltmann, Sara Lafia, Deborah L. Paul, Shelley A. James, David Bloom, Nelson Rios, Shari Ellis, Una Farrell, Jessica Uttrup, Michael Yost, Edward Davis, Rob Emery, Gary Motz, Julien Kimmig, Vaughn Shirey, Emily Sandall, Daniel Park, Christopher Tyrell, R. Sean Thackurdeen, Matthew Collins, Vincent O'Leary, Heather Prestridge, Christopher Evelyn, Ben Nyberg

Corresponding author: Katja C. Seltmann, Chedie Center for Biodiversity and Ecological Restoration, University of California - Santa Barbara, Santa Barbara, CA, United States of America

Articles by this author in: Crossref | PubMed | Google Scholar

Sara Lafia, Department of Geography, Center for Spatial Studies, University of California - Santa Barbara, Santa Barbara, CA, United States of America

Deborah L. Paul, Corresponding author, iDigBio, Gainesville, United States of America

Florida State University, Tallahassee, United States of America

Articles by this author in: Crossref | PubMed | Google Scholar

Abstract

Georeferencing is the process of aligning a text description of a geographic location with a spatial location based on a geographic coordinate system. Training aids are commonly created around the georeferencing process to disseminate community standards and ideas, guide accurate georeferencing, inform users about new tools, and help users evaluate existing geospatial data. The Georeferencing for Research Use (GRU) workshop was implemented as a training aid that focused on the creation and research use of geospatial

In standardized data to be published and shared, it is best practice to leave a field blank when no date (or other information) is available, rather than a placeholder.

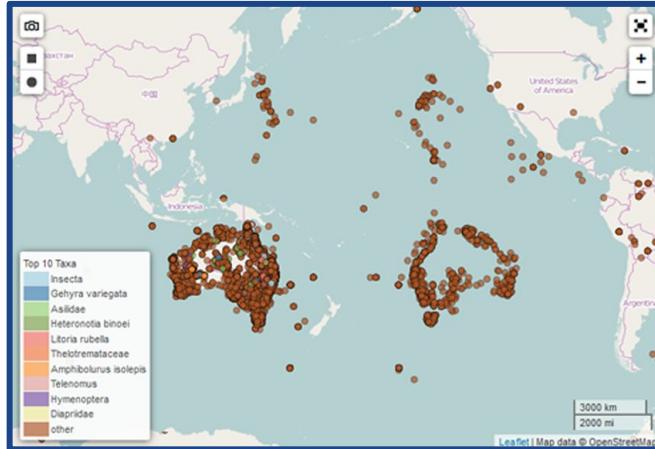


Grouping data issues

- time (date)
- locality
- place

 Hannah Frost
@feefifofannah Following

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



Country united k

united kindgom

united king

united kingdom

List ts

Family stitutiode

united kingdom (england)

united kingdom (scotland)

united kingdom (wales)

united kingdom [?]

united kingdom of great b

united kingdom?



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



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	<i>calosoma semilaeve</i>
1900-04-01	<i>platynus brunneomarginatus</i>
1903-04-03	<i>calathus ruficollis ruficollis</i>
1901-08-17	<i>cicindela trifasciata sigmoidea</i>
1901-08-17	<i>cicindela trifasciata sigmoidea</i>
	<i>cicindela senilis</i>
	<i>cicindela senilis</i>
1903-05-02	<i>omus tularensis</i>
1903-05-02	<i>omus californicus</i>
1903-05-02	<i>omus tularensis</i>



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 Ziemann¹, Yotam Eren^{1,2} and Assam El-Osta^{1,3*}

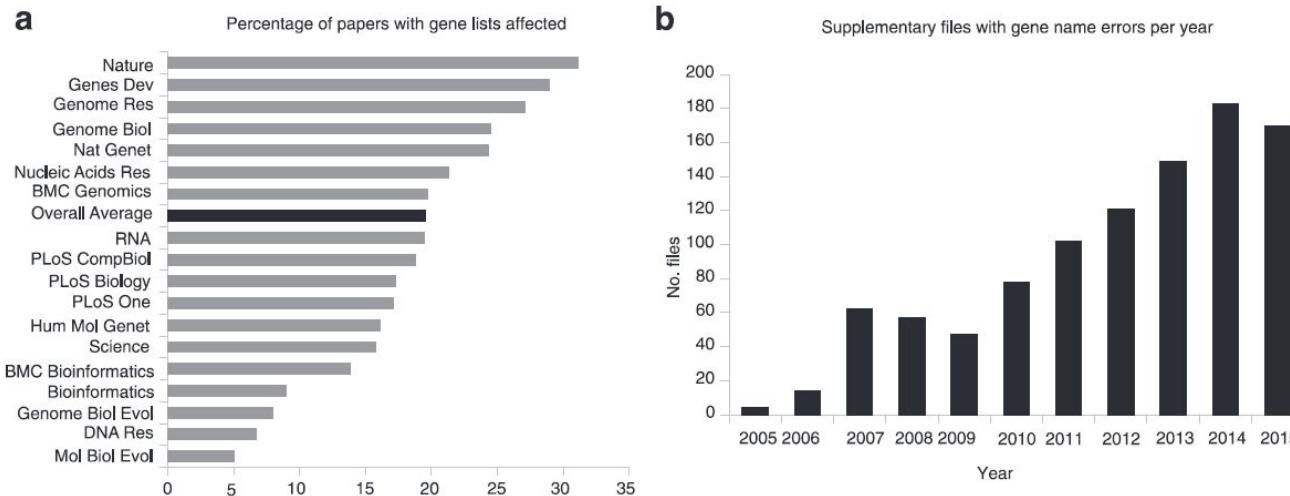
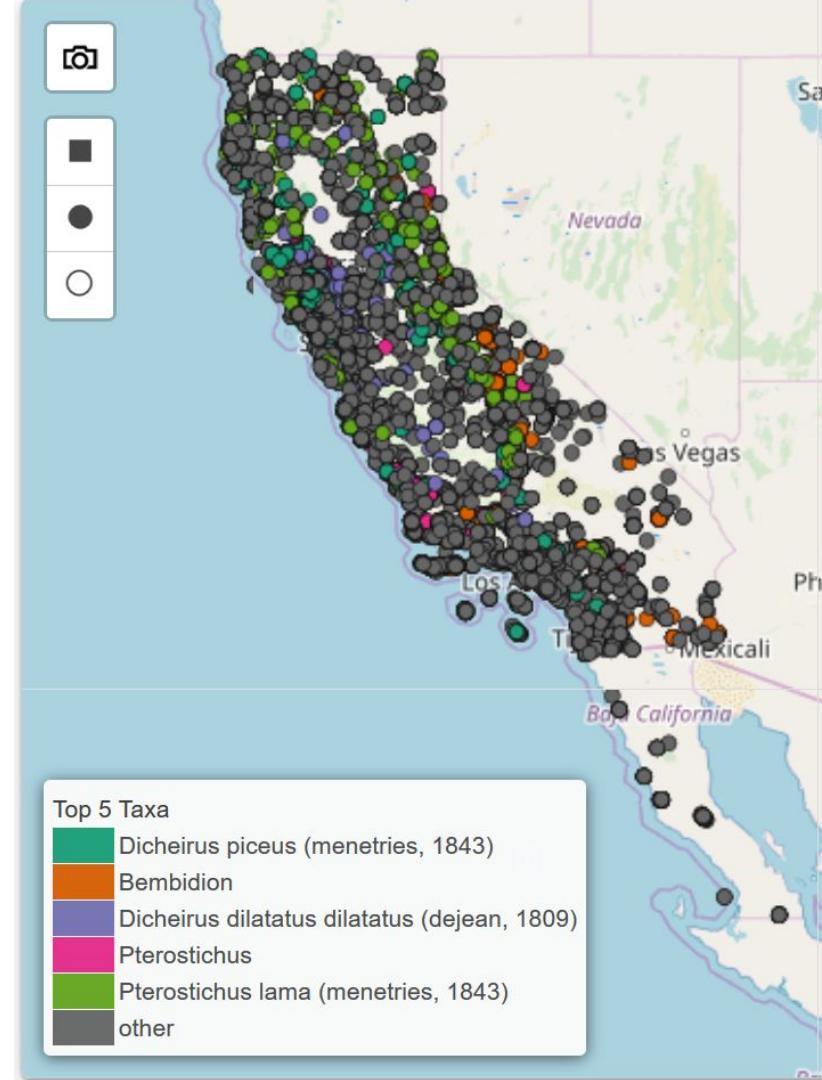


Fig. 1 Prevalence of gene name errors in supplementary Excel files. **a** Percentage of published papers with supplementary gene lists in Excel files affected by gene name errors. **b** Increase in gene name errors by year



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?

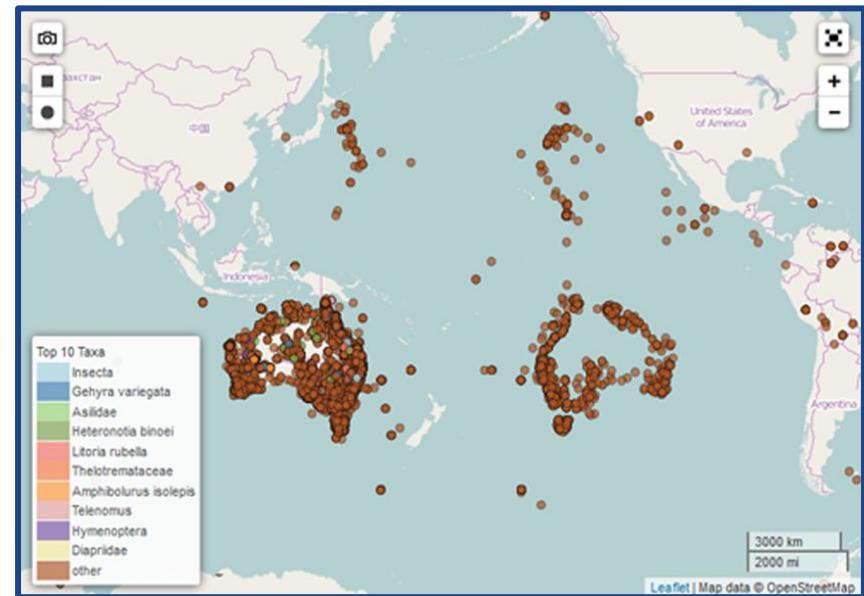
1. Remove records with undetermined, indet, empty
2. Use **dwc:taxonRank** to remove data outside ranks needed
3. Check raw and automated taxonomic names
4. Check endings - are they really the same?
5. Check synonymies, decide which names to use

CN	CO
:dwc:scientificName	dwc:taxonRank
calosoma semilaeve	species
platynus brunneomarginatus	species
calathus ruficollis ruficollis	subspecies
carabidae	family
cicindela trifasciata sigmaoidea	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

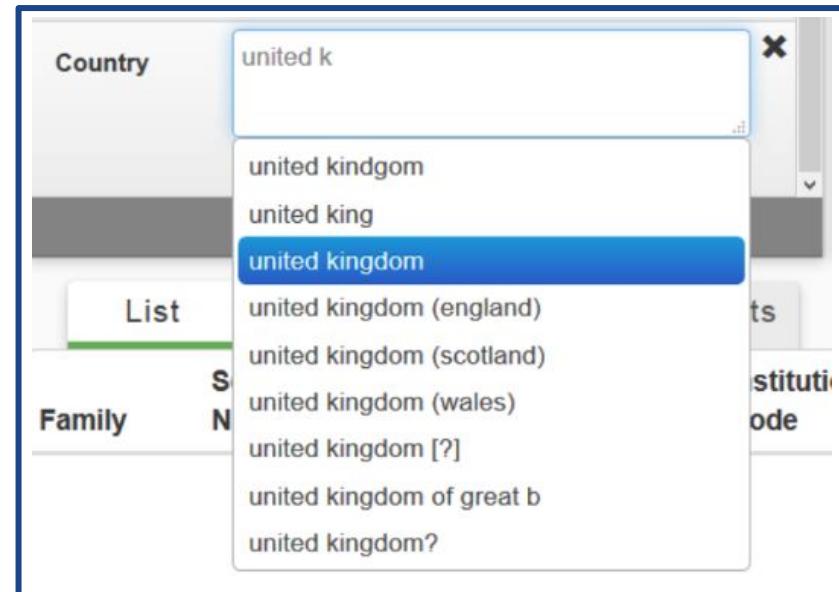
28°N, 80°W	YOU'RE PROBABLY DOING SOMETHING SPACE-RELATED
28.5°N, 80.6°W	YOU'RE POINTING OUT A SPECIFIC CITY
28.52°N, 80.68°W	YOU'RE POINTING OUT A NEIGHBORHOOD
28.523°N, 80.683°W	YOU'RE POINTING OUT A SPECIFIC SUBURBAN CUL-DE-SAC
28.5234°N, 80.6830°W	YOU'RE POINTING TO A PARTICULAR CORNER OF A HOUSE
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!"
28.523457182818284°N, 80.683094159265358°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



The screenshot shows a navigation bar with links for Home, Code of Conduct, Setup, Reference, Episodes, Extras, and License. A search bar is also present. Below the navigation is the title "QGIS Natural History Collection Data Lesson". A paragraph of text at the bottom states: "This tutorial was developed in support of QGIS training for a workshop in October 2016. The link to the workshop agenda is [here](#). This tutorial is for curations managers and researchers who would like to perform spatial data visualization and analysis using Geographic Information Systems (GIS)."



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



Amber



Christina



Christine



Deborah



Dmitry



Jeanette



Karl



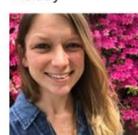
Kelsey



Kristin



Kyle



Laura



Margaret



Megan



Nico



Rebekah



Stevan



William

