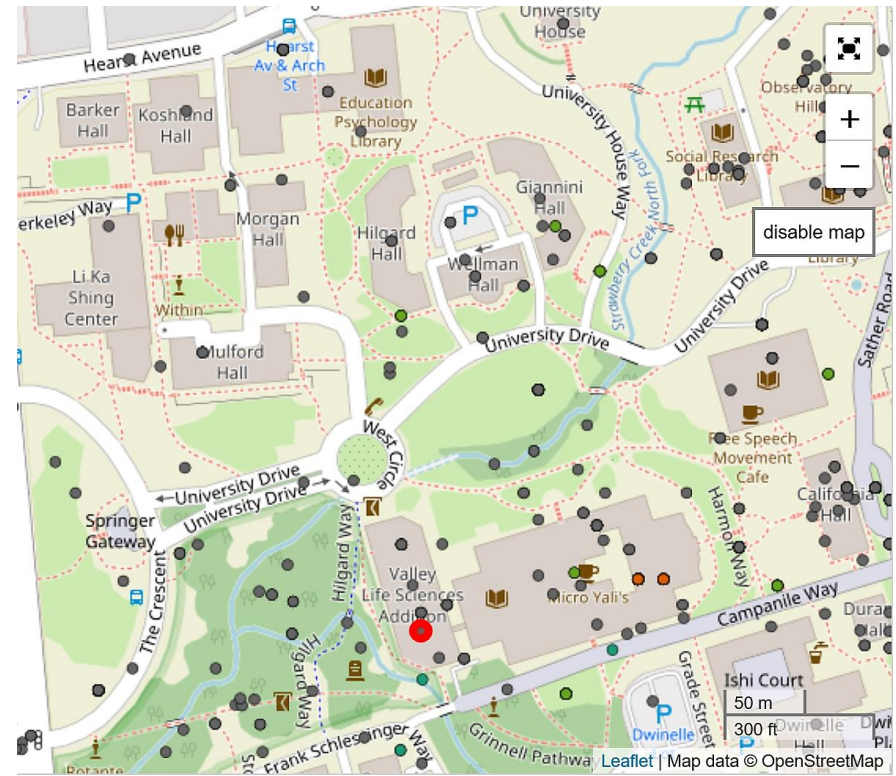




Learning from each other:
insights from discussions with
biodiversity data users and
creators about
what, when, and where.



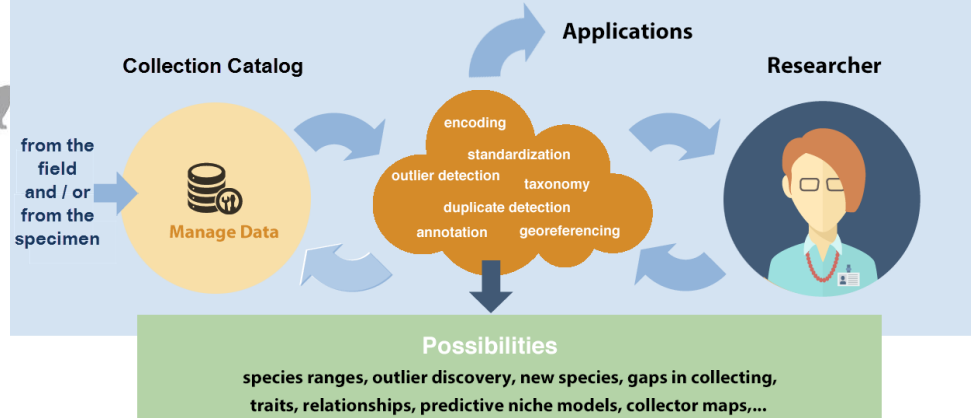
@idbdeb

Deborah L. Paul, Florida State University, @iDigBio
Digital Data II: Emerging Innovations for Biodiversity Data
Berkeley, California 2018



Georeferencing for Research Use – a community conversation

- ecologists
- systematists
- taxonomists
- conservation
- agriculture
- educators
- collection managers
- data managers
- software developers
- aggregators





overview of the workshop topics and discussions

- pre-workshop assignments
- 2 days georeferencing foundations
- 2 days focus on evaluating data-fitness for research use
- a mutual conversation



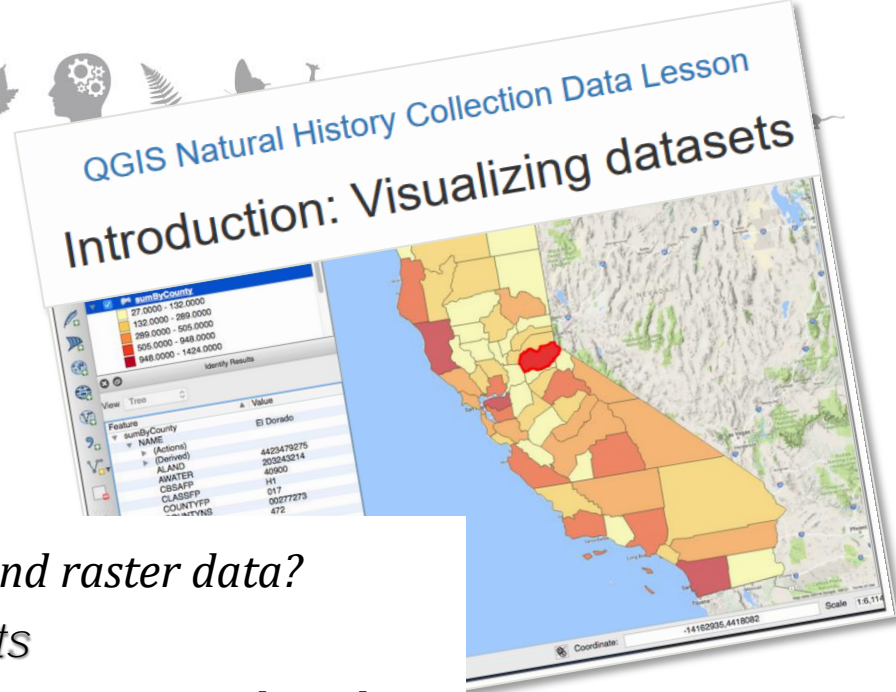


topics overview

- georeferencing basics
- data standards, terminology
- workflows
- field data collection
- data management
- legacy data georeferencing
- downloading datasets
- assessing data quality / uncertainty
- tools (R, QGIS, Open Refine, ...)
- creating maps
- spatial analyses
- hands-on bring your own data

bit.ly/GRUworkshop

Time	Activity
9:00	Review and Questions
9:05	<p>Georeferencing for Research Use Workshop - iDigBio Data</p> <ul style="list-style-type: none">• Downloading datasets from iDigBio - get data from portal, explain each component to the dataset. <p>filter and get the dataset</p> <ul style="list-style-type: none">• What is raw vs not raw?• Similar or different from GBIF?• List of iDigBio Flags:• Walk through steps of download, but provide dataset.• iDigBio Data set: http://s.idigbio.org/idigbio-downloads/a69d1541-4726-465d-84ad-50c7ed556eee.zip
	<p>Data Quality: How to evaluate existing georeferenced data/Fitn Use</p> <ul style="list-style-type: none">• Data quality flags• What do you... <p>Home</p> <h2>Data Organization in Spreadsheets for Natural History Collection Data</h2> <p>Good data organization is the foundation of any research project. Most researchers have data in spreadsheets, so it's the place that many research projects start.</p> <p>We organize data in spreadsheets in the ways that we as humans want to work with the data, but computers require that data be organized in particular ways. In order to use tools that make computation more efficient, such as programming languages like R or Python, we need to structure our data the way that computers need the data. Since this is where most research projects start, this is where we want to start too!</p> <p>In this lesson, you will learn:</p> <ul style="list-style-type: none">• Good data entry practices - formatting data tables in spreadsheets• How to avoid common formatting mistakes• Approaches for handling dates in spreadsheets



QGIS lesson set:

developed by UCSB graduate student
Sara Lafia

- *Introduction: Visualizing datasets*
 - *What is the difference between vector and raster data?*
- *Introduction: Preview and explore toolkits*
 - *What kinds of auxiliary data can complement spatial analysis?*
- *Exploration: Aggregating by regions*
 - *How can observations be aggregated by a given spatial unit of analysis?*
- *Exploration: Time animation*
 - *How can observations be checked for errors given dates or transcriptions?*
- *Exploration: Uncertainty*
 - *How can observations be symbolized to show systematic error and uncertainty by collector, or data quality score?*
- *Exploration: Next steps*
 - *How can observations be edited once errors are detected?*



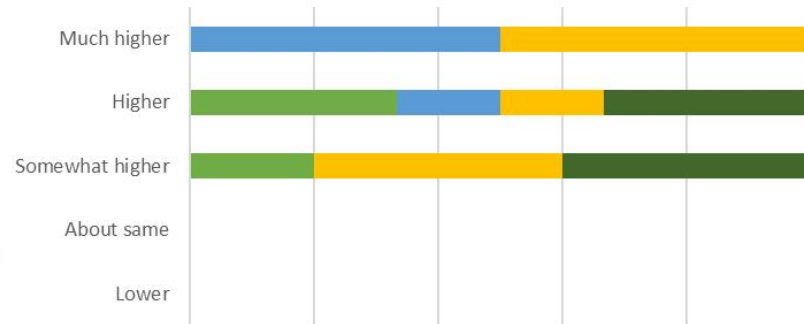
positive feedback



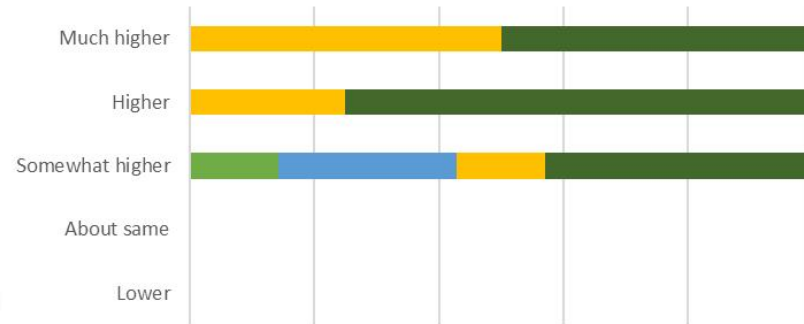
<https://www.leader-values.com/wordpress/nuance-what-does-buy-in-really-mean/> cc-by-nc-nd



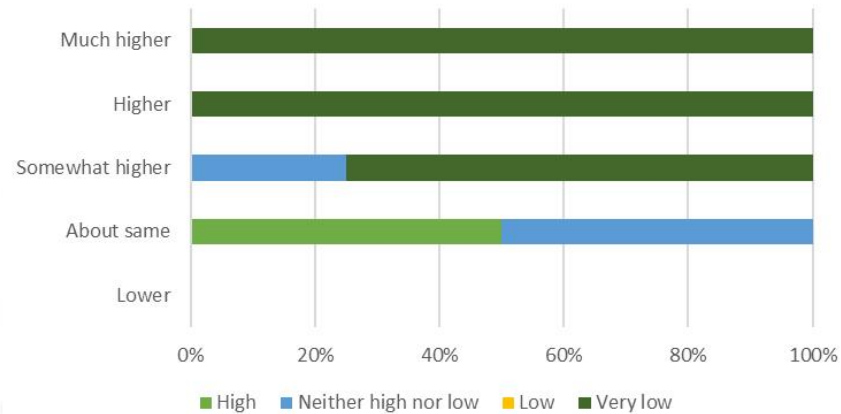
GEOLocate



OpenRefine

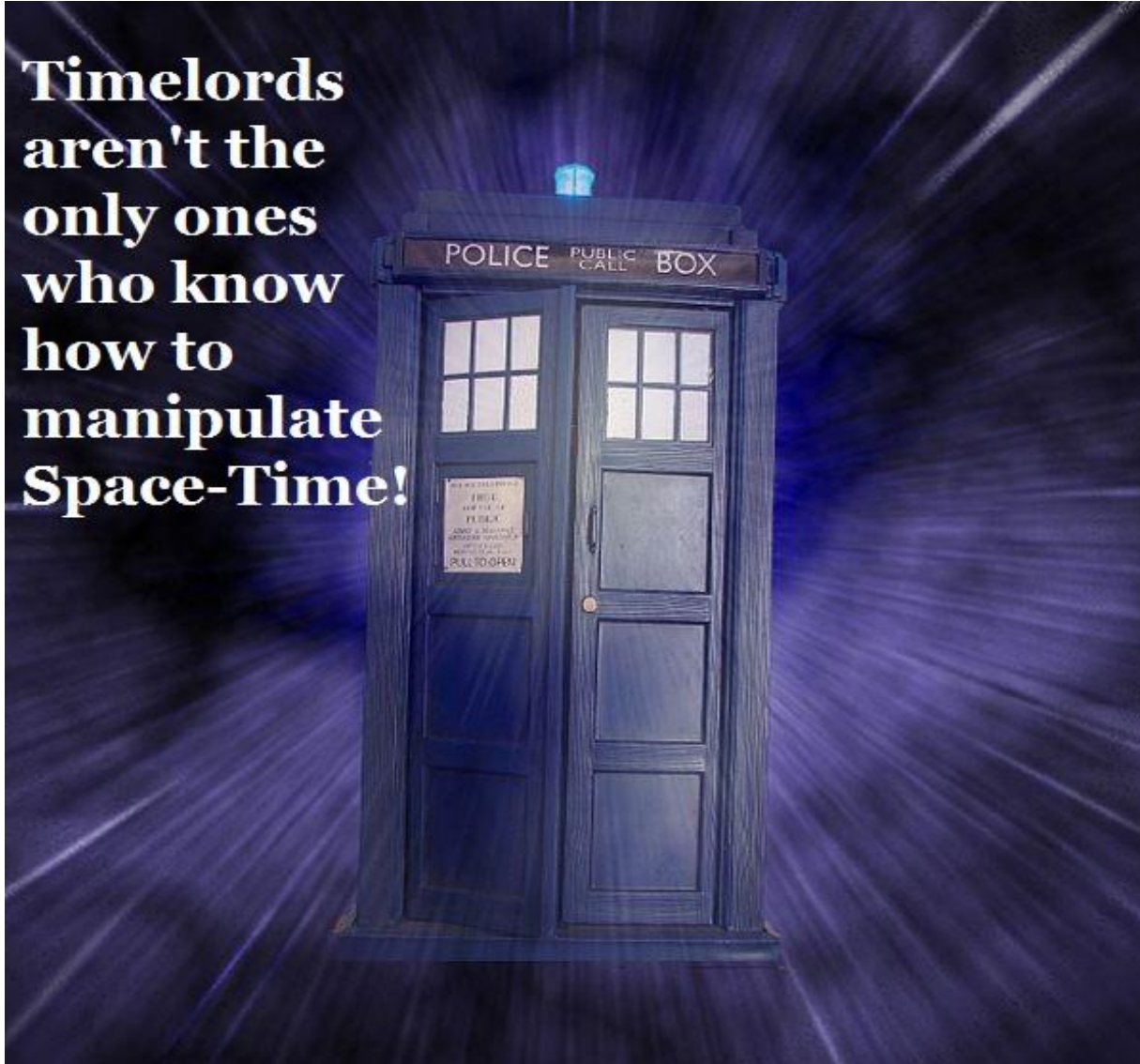


QGIS





**Timelords
aren't the
only ones
who know
how to
manipulate
Space-Time!**





data issues: time, location, authority files...

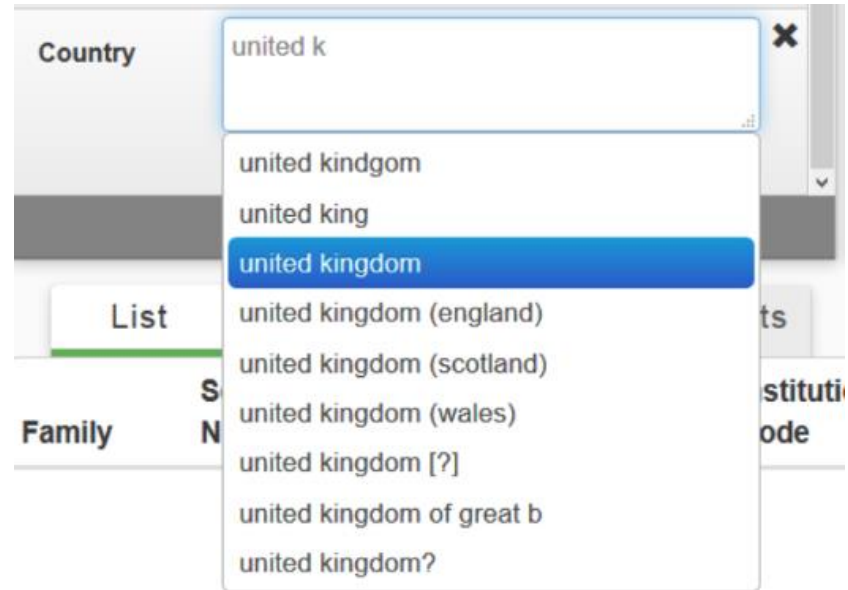
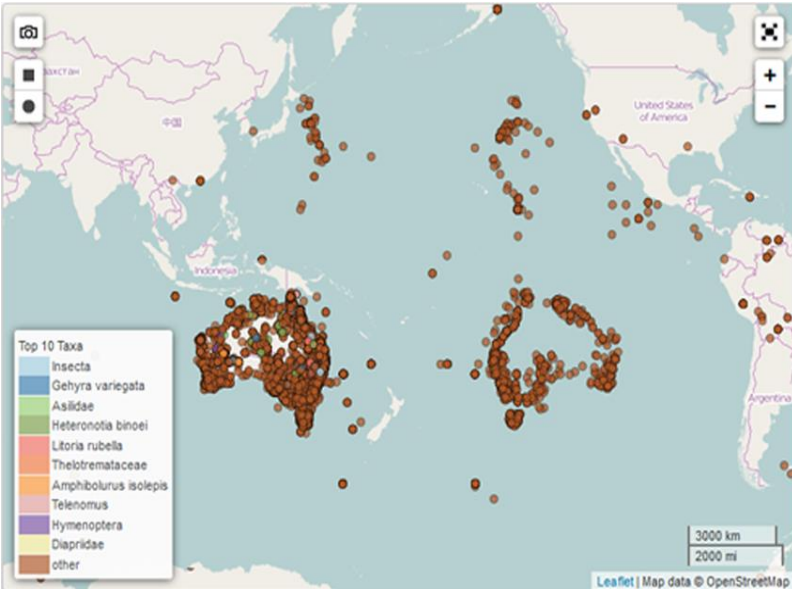


Hannah Frost
@feefifofannah



Following

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





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*
 - transparency please
 - concepts
 - authority files
 - parsing

GRU Workshop Conversation on Data Quality Considerations and Checks.

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

Workshop participants discussed data issues they look for and then generated a set of data quality checks to be considered when evaluating, cleaning, and improving data fitness for use. These were divided into three categories: time, geography, and taxonomy. Keep in mind this dq discussion focuses specifically on issues to look for in biocollections data. Addressing dq issues takes time and needs vary by research question. Each researcher will have to decide how much record cleaning (vs. record deletion) to do to best suit time constraints and the scientific questions. When evaluating and cleaning data – it is important to 1) save an untouched copy of the raw data – and 2) write down all steps taken when cleaning and standardizing the dataset (White 2013).

(annotation format: issue is listed, followed by a brief explanation to clarify some the dq observations, suggested tests and salient issues. The prefix “dwc” indicates a reference to a term in the [Darwin Core Standard](#) (dwc)).

Time:

- **Problems with 9999 dates** (or other **placeholder values researchers use to represent no data**). 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.



feedback from participants

- QGIS very helpful, found it much more accessible than ArcMap.
- Excited to go back and clean data!
- Interacting with different people and seeing how they georeference.
- Two aspects - validation of things I was already doing, and stuff I was struggling with - learning new techniques and tools. Now ready to sit down and write my own workflow.



uncertainty point-radius and polygons

GEOLocate Web Application
1 possible location found.

100 m
500 ft

bing
© 2018 Microsoft Corporation © 2018 HERE

Workbench 1 possible location found

Show 8 entries

Locality String	Country	State	County	WGS Lat	WGS Lon	Correction Status
St. Marks Nat'l Wildlife Refuge (Wakulla Unit), in shaded shallow water of wooded	United States	FL	Wakulla			
Frequent in rich sandy loam in mature hardwood forest on steep ravine slopes of W	United States	FL	Gadsden			
3.5 Miles S. Shoshone Amargosa River	United States	CA	Inyo	35.927353	-116.264549	yes

GEOLocate Web Application
1 possible location found.

100 m
500 ft

bing
© 2018 Microsoft Corporation © 2018 HERE

Workbench 1 possible location found

Show 8 entries

Locality String	Country	State	County	WGS Lat	WGS Lon	Correction Status
St. Marks Nat'l Wildlife Refuge (Wakulla Unit), in shaded shallow water of wooded	United States	FL	Wakulla			
Frequent in rich sandy loam in mature hardwood forest on steep ravine slopes of W	United States	FL	Gadsden			
3.5 Miles S. Shoshone Amargosa River	United States	CA	Inyo			



what's (needed) next

- born digital georeferences
- further tool development (GEOLocate, ...)
- locality services to reduce re-georeferencing
- publish and link georeferenced research data sets to the original occurrence records.

- more workshops around research use of the data requested by the participants (many topics)

- look for Pensoft RIO paper *coming soon*
 - *research use not included in TCN*
 - *workflows changing*

- your ideas...





Let's ~~be~~ clean up some space-time for future time travelers! Thanks, from the planet.



www.idigbio.org



facebook.com/iDigBio



twitter.com/iDigBio



vimeo.com/idigbio



idigbio.org/rss-feed.xml



<webcal://www.idigbio.org/events-calendar/export.ics>