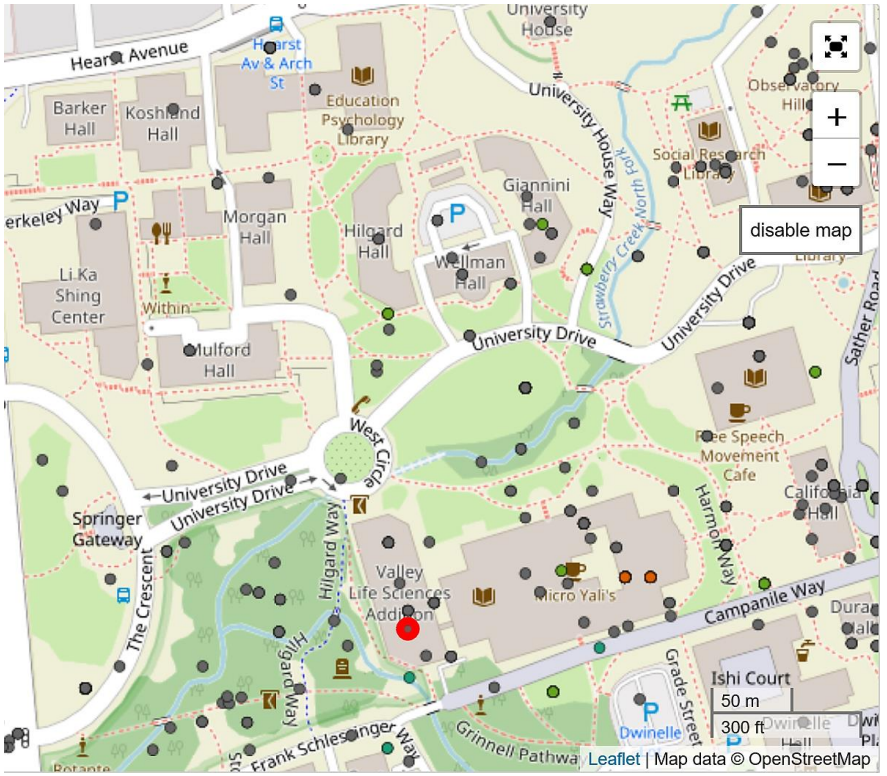
iDigBio
Integrated Digitized Biocollections

Learning from each other:

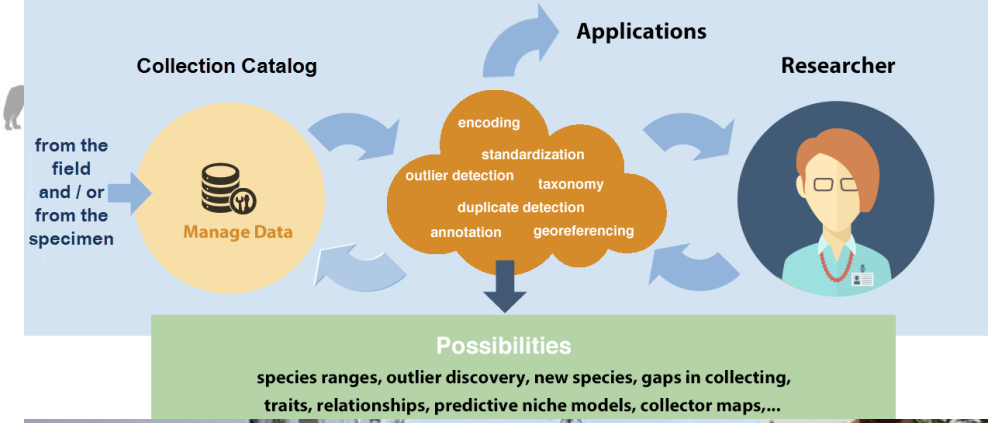
insights from discussions with biodiversity data users and creators about *what, when, and where*.



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

iDigBio is funded by grants from the National Science Foundation's Advancing Digitization of Biodiversity Collections Program. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

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

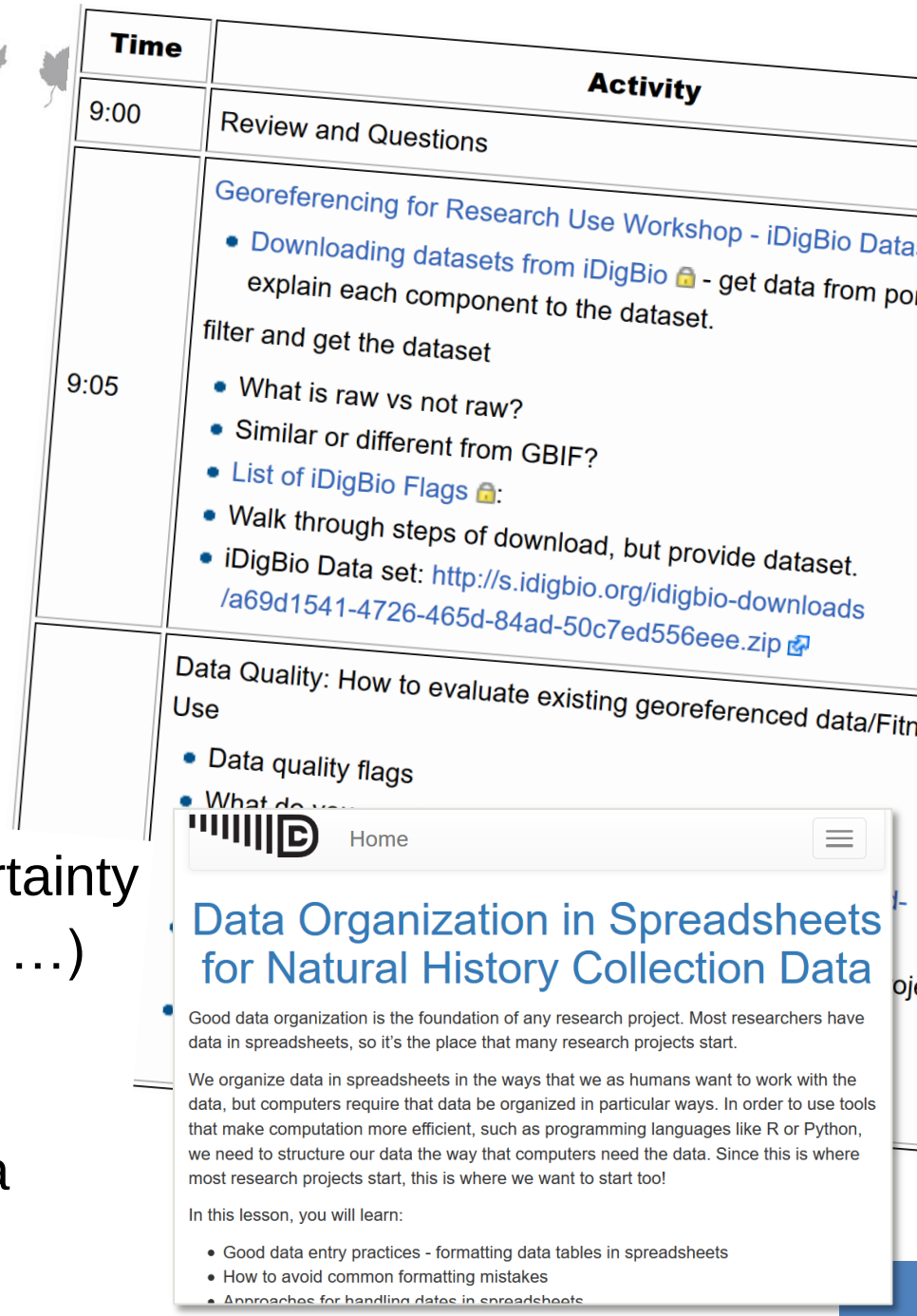
- a mutual conversation

http://ironyfunny.tistory.com/206

topics overview

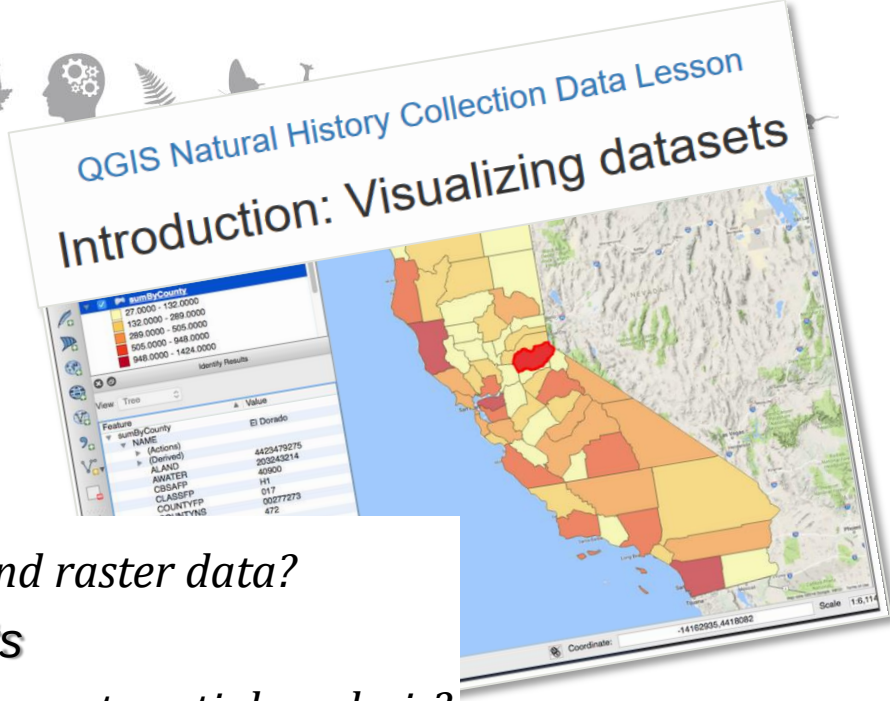
- georeferencing basics
- data standards, terminology
- workflows
- field data collection
- data management
- legacy data georeferencing
- downloading datasets
- assessing data quality / uncertainty
- tools (R, QGIS, Open Refine, …)
- creating maps
- spatial analyses
- hands-on bring your own data

bit.ly/GRUworkshop



QGIS lesson set:

developed by UCSB graduate student Sara Lafia

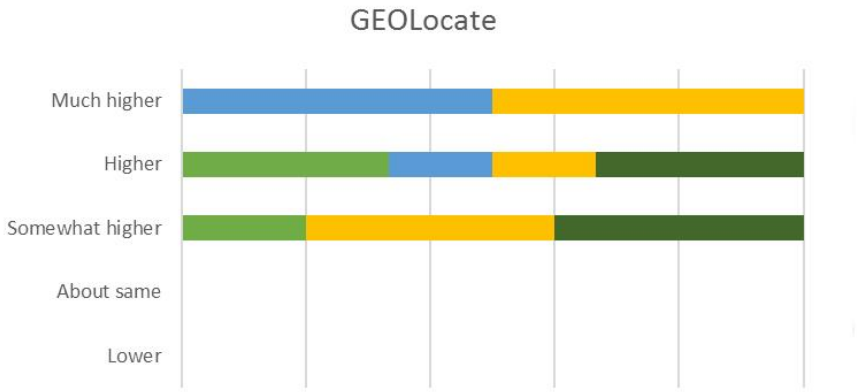


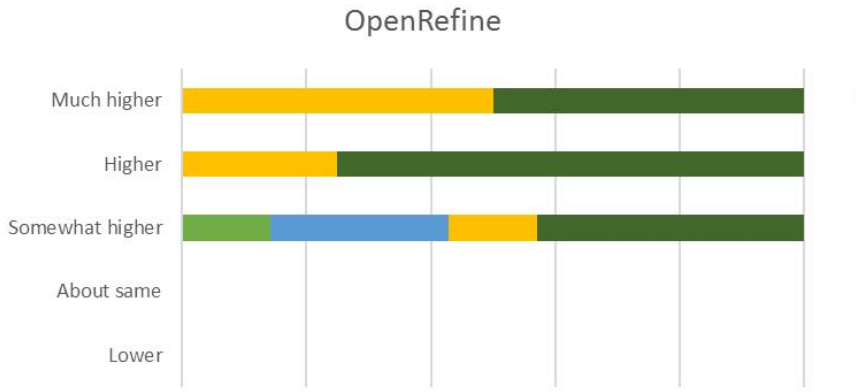
- *Introduction: Visualizing datasets*
  - *What is the difference between vector and raster data?*
- *Introduction: Preview and explore toolkits*
  - *What kinds of auxiliary data can complement spatial analysis?*
- *Exploration: Aggregating by regions*
  - *How can observations be aggregated by a given 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?*

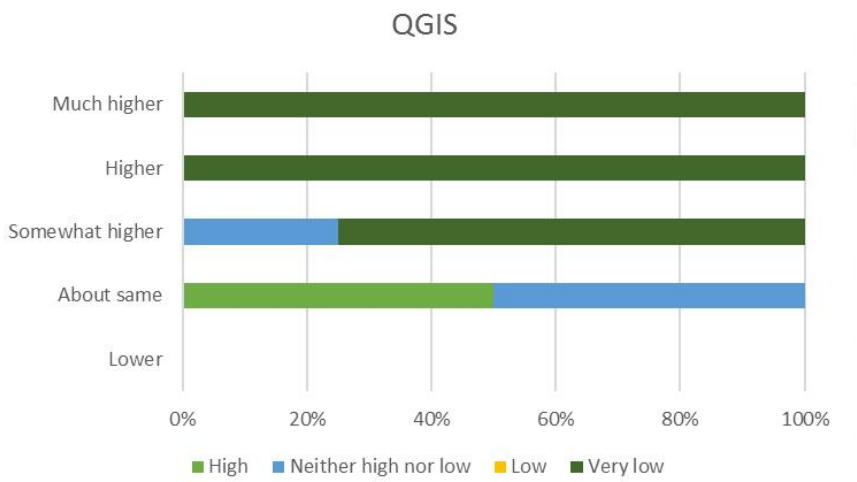
https://data-lessons.github.io/QGIS-nhcdata-lesson/

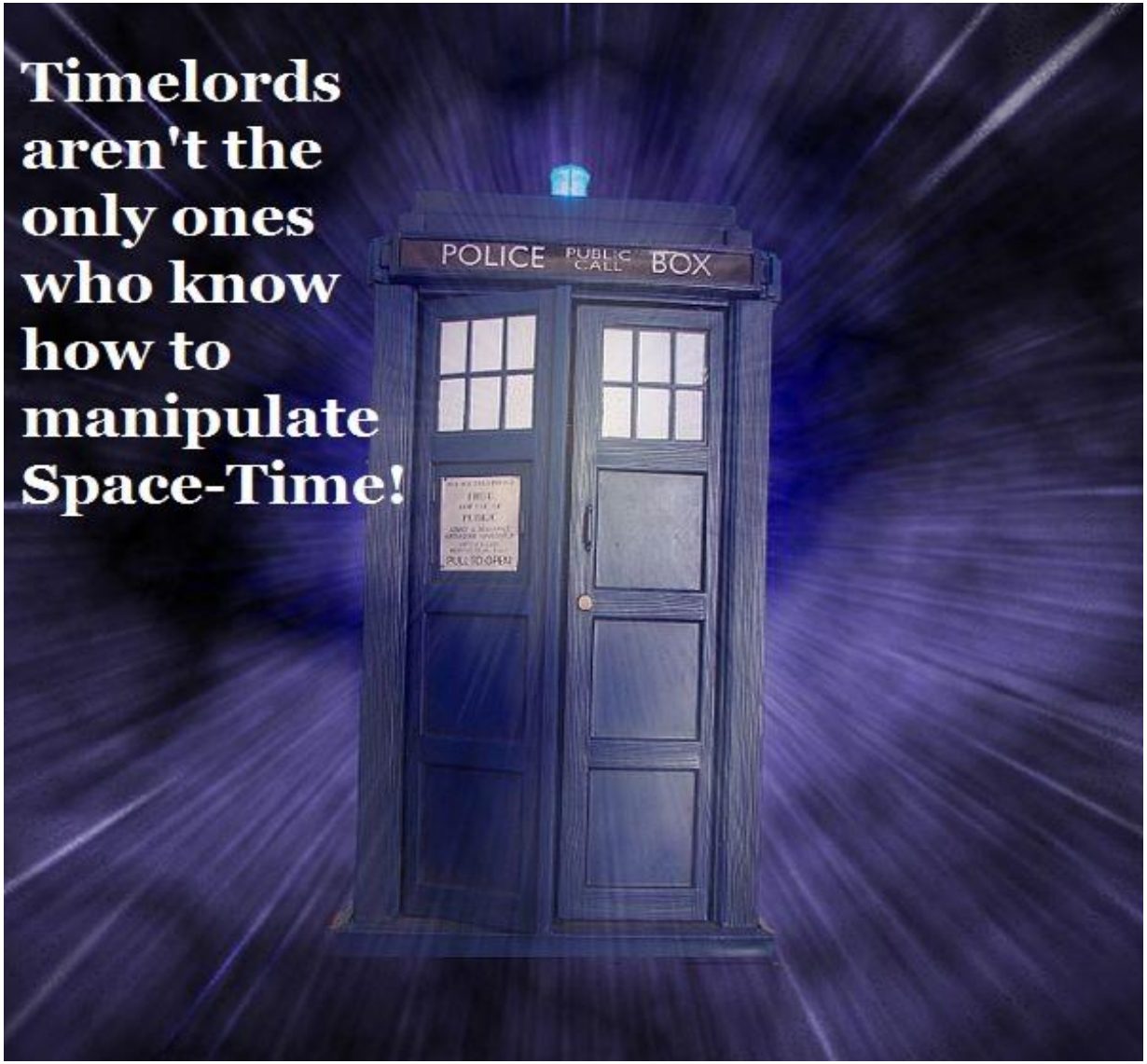
positive feedback

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

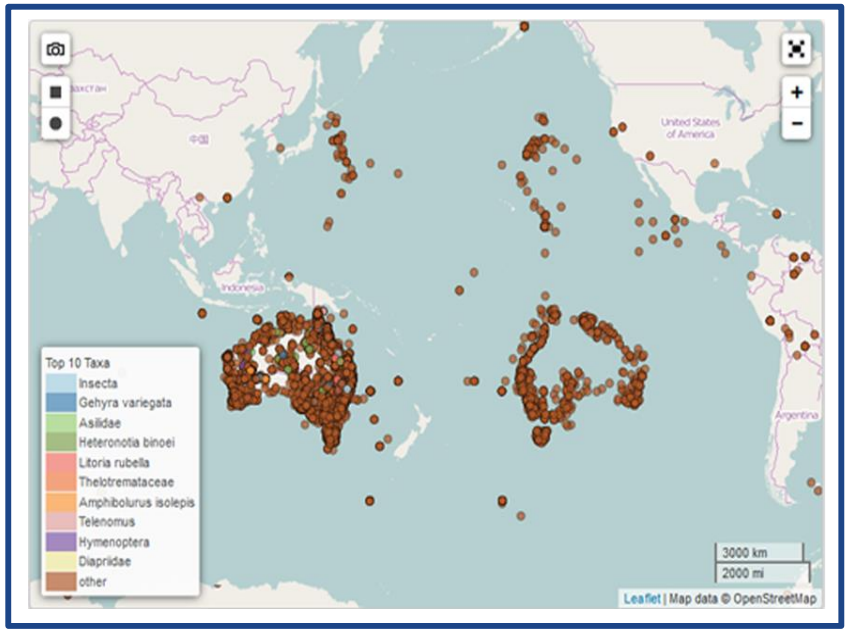


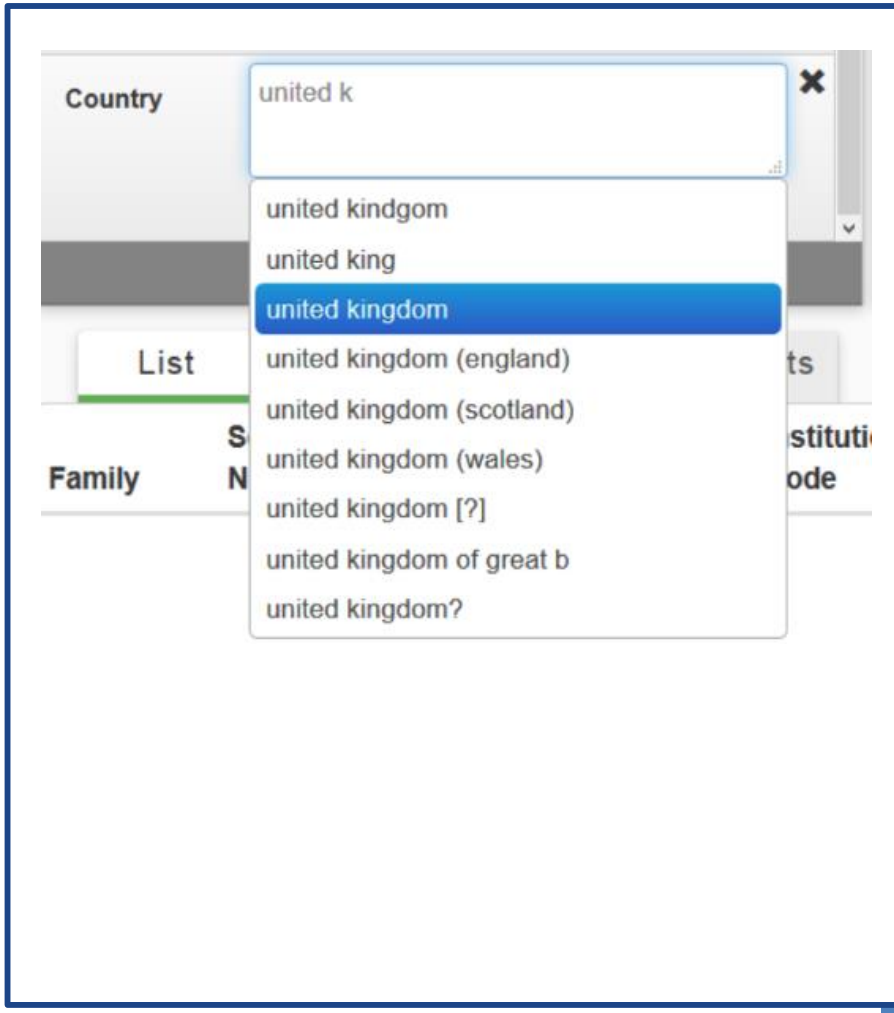





Timelords aren't the only ones who know how to manipulate Space-Time!
POLICE PUBLIC CALL BOX

data issues: time, location, authority files…

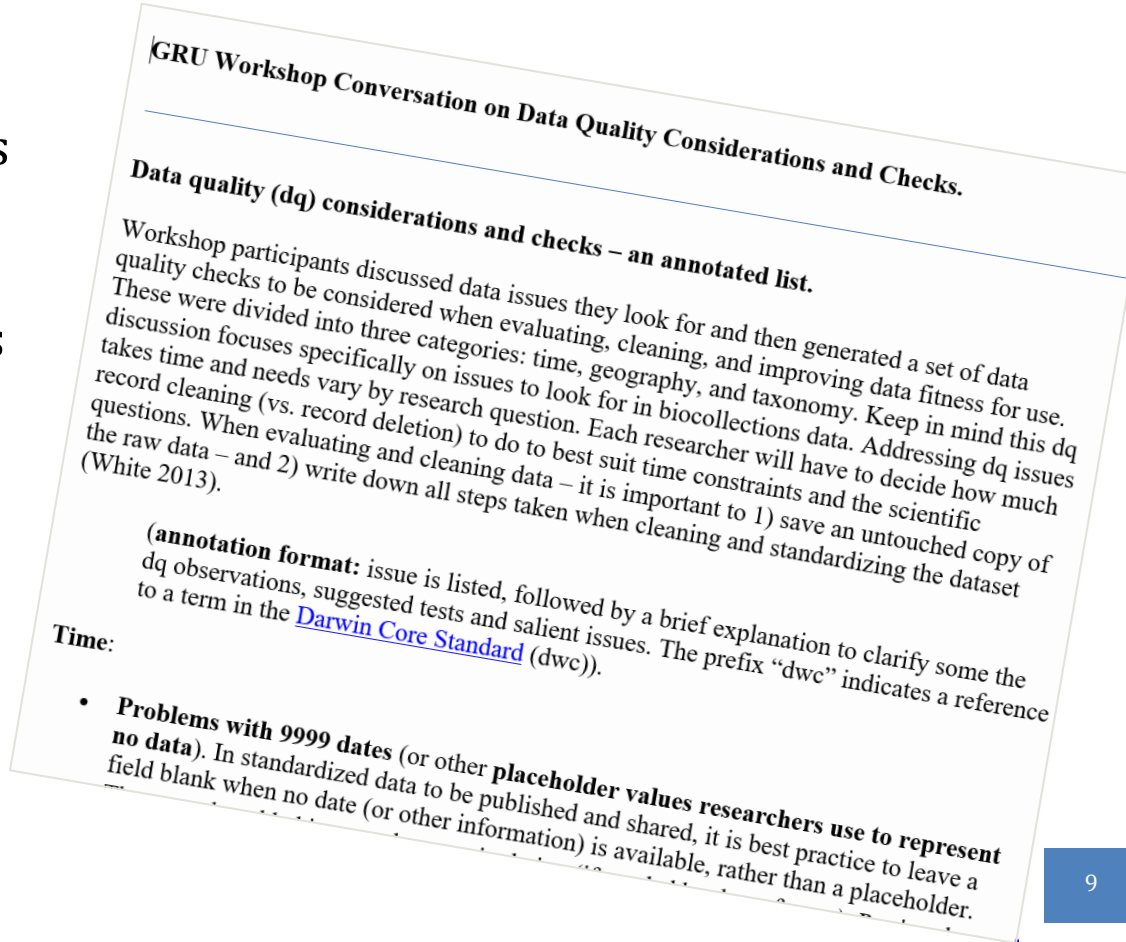




synthesis of issues:

-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)).

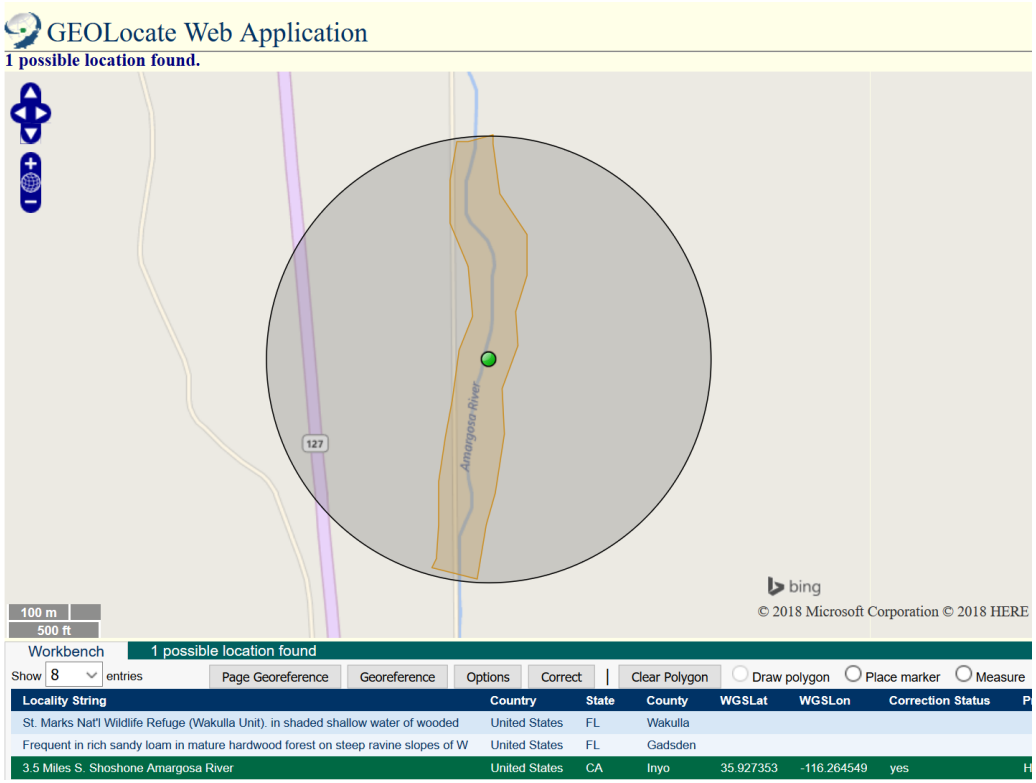
Time:

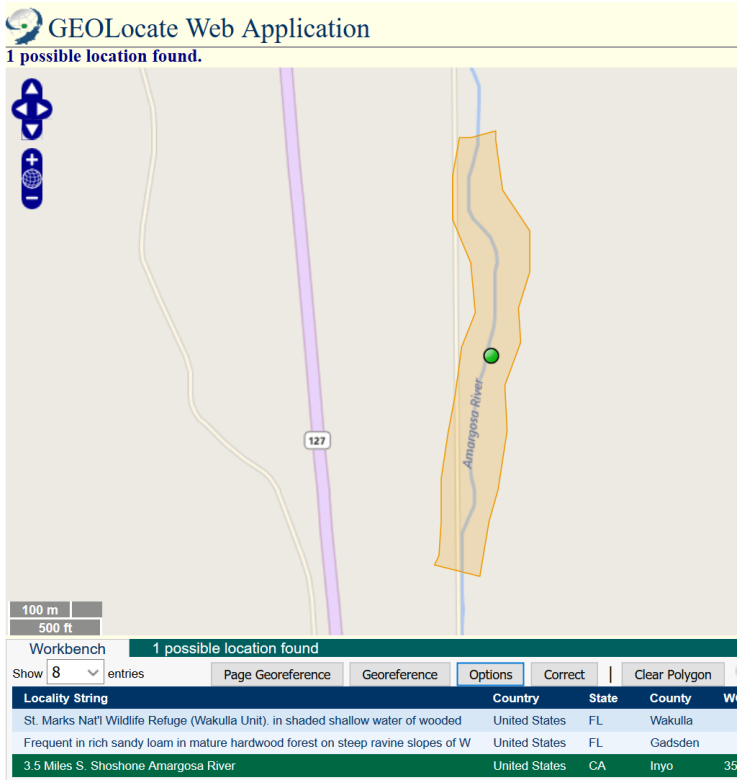
- **Problems with 9999 dates** (or other **placeholder values researchers use to represent no data**). In standardized data to be published and shared, it is best practice to leave a field blank when no date (or other information) is available, rather than a placeholder.

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





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 ~~bend~~ 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

iDigBio is funded by grants from the National Science Foundation's Advancing Digitization of Biodiversity Collections Program. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.