iDigBio Summit IV : Data Management

Monday, October 27, 2014, 1:30pm - 2:30pm

Organizers: Larry Page

Potential topics include feedback and annotations to data, taxonomic names, services including searching by higher taxa and synonyms, extending data schemas to include additional (non-Darwin Core) information, and attribution

Discussion Points

Type your notes here.

Attribution (getting credit for use)
Notification of use of data
Data transformation
Extending data schemas to include additional
Data products: transformation from users

Is there a responsibility for data preservation?

Kurator Project
- U of Champaign-Urbana
- successor to filtered push
- means of returning annotations, corrections to update records
- curator built into data management software
- pipelines to check authority files to make sure data is correct
- name checked against authority file, if record not on list, would be flagged - misspelling, or recent paper coming out with new name
- means of identifying mistakes in the data
- compare georeferenced location with natural distribution of a given species, if outlying the distribution then suspected of being incorrect, is flagged
- ability to generate annotation, with choice to accept or reject, embed comment, whether you accept or reject

all under assumption that data could be put back into the source database

Mistakes could be flagged and sent back to data owner: data owner can accept or reject, if accept record gets automatically changed
Different names for this process:

- annotation (versioning); bring in data from other portals; message goes back to the source to get corrected, keep track of the versions

- Symbiota: version table, edits applied by approved editor (table can be added to Darwin Core archive)
- iDigBio - history of record deleted from system when updated (versioning)
- sensitive species: removing images, redact records if they are sensitive

Attribution:

- what level of attribution?
- with a data set there is is an inherence of attribution
- interpret all attribution and up to user to decide
- can you copyright a record?
- scholarly attributions
- biggest issue - obtaining appropriate credit
- training is essential for attribution - we all need to be teachers
- efforts to make it easier for people who do not know how to cite properly? could be created using an algorithm
- GBIF led the way: attribution block, iDigBio following same format - should be able to copy and paste into manuscript, etc.
- links to NORMS (http://www.vertnet.org/resources/norms.html), each institution when published, citation for each record when downloaded, or full blown citation for whole data set

How is use of data going back to the provider?

- how to enforce?
- information is starting to be collected, difficult to summarize all uses of data
- lacking: API tracking usage, important to have a unified tracking mechanism
- on screen use and downloaded use
- Example: http://www.vertnet.org/resources/usagereportingguide.html

Best practices webinar may be needed here

If available, would love to use the information, because of reporting to NSF

- improve profile to administration

Great to have all aggregators report similar kind of metrics

Possibly want on a schedule or self serve?

- self serve
- should not be a lot of work to make available

Data transformations

- means something different to everyone present
- iDigBio - working on developing workloads and dataset to make more searchable

- most advanced: pull #s out of elevation and depth fields, to pull out into actual number that can be searched
- validating lat and longs, cleaning states and countries, really hampers use of data
- moving from raw data products to authority files
- how to get information back to the provider?
- not finalized yet on iDigBio side
- no way to push data back into collections, even if you can hand to the provider, they can't do anything with it
- secondary vs. primary data
- want to provide data that is useful
- additional setup for FilterPush
- implementing FilteredPush at higher level
- iDigBio focusing on correction (e.g., Florida)
- API with a list of changes

Work on appliance that automatically employs Specify where FilteredPush would be default
Data transformations and annotations - sense from people on ground - how to enrich data that we already have; record sets already ingested, when trying to answer questions, needs an iteration of data