# DATA DISCUSSION GROUP

Facilitator: Gil Nelson

Location: Cypress

<https://docs.google.com/document/d/1nGvDSLMjn4xwJDQ17rnpODpd7EvxJuC5xK6Q55VKKAQ/edit>

[**http://tinyurl.com/summit2-data**](http://tinyurl.com/summit2-data)

**Notes by Charlotte Germain-AUbrey**

**Alex: Globally unique identifier. Is it tied to the specimen itself ? could be based on the barcode, like is done in GBIF. iDigBio**

**Toby: some have tied GUID tied to taxon names, but not a good idea. Place in phylum, family, and work down the tree and assess identifiers to nodes in the tree.**

**Alex: does not make sense conceptually. Better to use a name identifier from TROPICOS.**

**Katja: What happens when the original source changes the identifier or name ?**

**Paul: GBIF checklist bank. What is the ID of this family ? Not clear if it’s a ref to the nomenclature or a taxonomic concept.**

**Alex: ID identifies a string ?**

**Paul: no. THe name string was created by this person in this publication and some third party says it’s a valid.**

**Taxon concept: in this monogram, this author put characters as defining a set of specimens. In insects, species number that people use.**

**Nomenclature: not as well defined.**

**Toby asks about the use of taxon concepts in TCNs.**

**disucssion about what people use. General concensus that in large collections, in order to verify the ID, keep track of the people who have identified the specimen, and leave it at that. Problem that no trace of what the identification was done with (what flora, monograph...). Impossible to revise that.**

**Alan weekley has done a lot on taxonomic concepts in plants. Emphasizes the need of annotating the source of the taxon concept at**

**Paul: groups trying to work out the informatics for taxon concepts. Rich Pannel, Bloom, Pattison work on that. Pieces of infrastructure are starting to emerge.**

**Alex: Nomenclatural act of all the TCNs, then identify local authorities, adn link them. Global database won’t make sense without transmitting the nomenclatural information from the local database.**

**Paul: list of plants from Patrick Sweeney of Northeast.**

**Patrick: hierarchy of name is separate from the identifier of the name itself. Independent of concept too.**

**It would make sense to have the GUID attached to a name with an authority. Name can be attached to ITIS but even ITIS is not updated for everything. Institutions do not want to become a name source.**

**Alex: in summary, names on specimens. Data provider feel comfortable with that name. the string is all you get as an identifier. NOT comfortable to assign a GUID to an identification history. T**he binomal is attached to the specimen itself, and potentially authority files (if collection manager provides a list of valid names, that is good enough to disseminate to other collections).

Ed: that doesn’t work for herbaria as researchers are the ones identifying and annotating specimens.

tanya: But manager needs to provide the names in the end because they might not appear anywhere else.

Alex: not everyone agrees on the same name for the same thing, but a collection manager agrees that one name means one thing (might be different from another manager).

Katja: problem at the community level. Why not push the names to another authority to assign those GUIDs to names ?

Alex and Gil: Does iDigBio need to take that role ? But then the authority is only for the name only (spelling, etc...), not for the quality of the assignment of the name to a specific specimen. Not an authority file itself, more of a name checklist.

Paul: same direction that GBIF took.

Gil: warning about iDigBio not having the role of the authority of the name itself, or the nature or quality that the managers are providing them. Need to stay focused on mission of digitizing specimens and making them available.

Alex: data provider can choose to give a checklist or not. Toby agrees that the quality of the data give to iDigBio will vary greatly according to collection manager, group having been revised recently or not, etc.... and we just have to accept that.

Omar: asks about iDigBio creating the GUID itself. Alex says that iDigBio does, but not expecting people to use that, or follow that.

Omar: how to trigger the creation of the identifier ? If I have a GUID and a map, we should coordinate.

Gil: does not work for smaller institutions. Getting the GUIDs back into the small dataset of a small collection and keeping track of them is not realistic.

Rod: suggests that some collections are going to have to be left behind. If they are not willing to use a prepackaged software. Need to motivate smaller collections to stop using excel and start using software that is more modern.

Alex agrees that we need to help them move forward and migrate those data to better software. Gil is concerned about leaving smaller collections behind.

Is it iDigBio’s responsibility to migrate those datasets ?

Toby thinks that it is iDigBio’s responsibility to host 3-4 softwares on the cloud that others can use. Symbiota is living on the iDigBio server, so this has started. Letting people use our servers solves the problem.

Alex thinks that iDigBio does not have the staff to maintain the software, but can provide the infrastructure necessary for others to use. Toby points out that some software are advanced enough to serve the purpose of smaller collections, so no need to actually create new code.

Jose emits the idea of launching a series of workshops for smaller collection managers with a certificate of competency.

Lucille has been on the waiting list to migrate her collection information to Specify for 4 years.

Problem of the middle man between small collection that are not even in Excel yet and the developers that improve the already advanced software.

Katja: idea of a migration hack-a-thon where smaller collections come with their datasets and others migrate their data into more modern software.

Rod: NSF is not providing the money necessary to receive all this data in Specify.

Gil and Jose talk about workshops by iDigBio. Jose: do a train the trainer workshop to get the projects starting.

Alex: don’t train the trainer but train the migrator.

Rod: no need to use Specify or even Symbiota, can train them to have a well-formed spreadsheet so iDigBio can ingest it.

Austin: have the applicants identify 3 (for example) collection databases/Excel spreadsheets that they will take into an up-to-date management system.

Toby: problem of institutions don’t have the money or staff to buy and maintain a server. If we could provide those servers and maintain them, then the first hurdle would be erased.

Alex: can use resources from others, it’s much cheaper. Need to coordinate the softwares that are being used though.

Jose: train the trainer workshop would also include the training to create and store GUIDs.

Alex argues that you need people to create their own GUIDs.

Paul: database managers have to manage their own GUIDs.

Alex: if one collection absorbs another one, you’re going to get new GUIDs, most likely different from yours. End up with multiple GUIDs for the same thing.

Paul says one GUID per object is the max.

Joanna: iDigBio agreed that things we accept have to have a GUID.

Katja: why assign a GUID to everything (geographical location for example) ?

Gil: Should we require some unique media object names ? GUIDs for filenames ?

Jose: we can work with everything that gets submitted to iDigBio has a unique ID. However, if something comes to us with a name that might not be unique, it’s a problem. iDigBio will assign a GUID to those objects.

Paul: what if we have a GUID generating tool and request people to assign one to each of the objects that they submit to iDigBio.

Jose: should iDigBio identify an approach that would deal with the GUID assignment and train the smaller collections: adopt new software, converting existing media into some other software, assigning GUIDs...

Katja suggests training the grad students involved into these collections.

Toby: need to include other managers in the workshops so they can talk to them about their own experience and why it is important, not only include technical people. The point about the middle people being important.

Joanna: migrating your data to another medium includes a cleaning process that might be long and a hurdle.

Gil: pre-digitization step very important.

Rod: Specify has explanations on their website about how to do these things.

Omar: Experience with disposable computing (Asia). Raspberry PI can run a web-based app ($4-5 flashcard) can be distributed to smaller collections to use. Hack-a-thon could use this concept. Video could be built-in for people to watch and get trained and migrate their data with this $35 disposable device.

Paul: potential issues with institutions only using certain machines, being under certain contracts, etc...

Rod: specify can work with PC, Mac and has a desktop version on a USB key that can take data from excel spreadsheets. Need to format things a certain way, which would be an important step to teach in the workshops.

Jose enquiring about the number of institutions that host collections: 160 institutions included in TCN, 1500 collections estimated in the US.

Andy: what are the co-PIs not doing to capture the smaller collection ?

Rod: 2 categories: willing to change the way they do things, or not willing at all.

Gil: vast majority of collection managers would like to upgrade their collections to better databases that can be shared, but not enough funds, time, knowledge, staff...

Gil: 450 collections have Specify, most of these collections small. don’t necessarily use it because uncomfortable, so need workshops to train these people and assist them.

Andy: smaller collections have to be picked up by TCNs to be able to capture, especially certain groups (bryophytes and lichens for example).

BREAK

Gil: on the iDigBio website, people will want to search higher taxonomic level, how can we include records with only a binomial.

Toby: family name is most important, unique.

Gil: MISC committee wondering how to most effectively convey information.

Toby: genus sp. for large genera: lost of records like that. how to categorize sp. ? Can the portal go past the blackbox of not having the full resolution for a specimen ?

Gil, Alex, Lucille, Toby: if you have added info, you can still get to this record: geography, habitat, etc...

Gil: if you search for a genus that has been put in different families, should be able to go to Tropicos and search all the families listed there and look for this genus in the records.

Toby, Alex: if you have required higher taxon rank, that would help with homonyms.

Patrick: assign something to a family, or multiple family.

Alex: global distribution of taxon classification. Need to be able to assign GUIDs to several Darwin Core domains to be able to search more efficienty and retrieve the redistributed.

I GOT DISCONNECTED AND MIGHT HAVE MISSED 5 MINS OF DISCUSSIOn.

problem of efficiency of search. Need some external authority (ITIS, TROPICOS, iPlant...).

GBIF requires one higher order, GCC (?) requires all higher ranks.

Patrick: we could require people to supply the family level, or parent to the family.

Katja: use index search of names of things and were it was taxonomically. Then, all possible parents of a Quercus search will be searched.

Toby: example of Mexico, taxonomy is simpler. Maintain lists of North America, restrain geographically to not have to deal with everything globally.

Toby: when you have giant searches, everything coming back has high interest.

Alex: improving system to make sure you can refine search to get results of interest.

Jose: discuss concrete things that iDigBio can implement.

Alex: cyberinfrastructure working group will start tool- or discipline-focused groups to see what and how data people can transmit. Need for everyone to think of what data is of value for you and for the broader community. Checklists might have more value than we think today. How can you identify a discrete checklist, from your dataset. All the way to species level ?

What can you use or modify to assign these persistent identifiers for the object.

hoping to get Specify and Symbiota to integrate this data (?).

Jose: 2 concerns:

1) if we ingest data and make it available, it has to accurately reflect what you submit (ie: how does a taxon fit into a category?)

2) fitness for reuse. can we do things with the data that would normally not be possible without iDigBio ? try to anticipate how your data could be reused, and you can use other people’s data.

**main points:**

Names still a very important and expansive issue

idigbio’s role is not to solve the namesAuthority files vs. checklists

robust searching when taxonomic data is lacking

index of related names using name services

may be a role for idigbio hosting checklists and name matching

importance of GUID for records in db-esp

small collections:

where should ids originate?

 provider

 idigbio

should idigbio provide workshop for GUID development

and to migrate data to more sophisticated db system

hackathon to assist in this migration; cleaning code, etc.

Toby:larger question beyond taxon name, e.g narrowing identification

idigbio initiatives:

via CI WG

1- what data can and want to submit?

2- data of value

3- how to create and send checklist or authority file

4- how to define persistent identifiers

5- idigbio to focus on existing data models: Specify, Symbiota, PBI

6- want to accurately reflect data as submitted and as contributors desire

7- fitness for use

8- how do you want to use data