12-03-16 BIEN db

Participants: Brad Boyle, Aaron Marcuse-Kubitza, Bob Peet, Martha Narro

Unavailable: Jim Regetz, Mark Schildhauer, Mike Lee

Agenda

Loading DwC inputs (specimen data)

Revising VegX XML schema

Aaron brought Bob Peet up to speed on what had been done over the past few weeks.

**VegX Schema**

* BP: At some point we’ll need to talk with Nick about changing the VegX draft standard and writing an addendum to the publication.
* BB: In VegX, aggregate and individual observations would be linked through something. The question is what. Link them through a user-defined field?
  + Brad likes Aaron’s solution that he came up with for VegBIEN.
  + Nick doesn’t want to make individual observation a child of aggregate observation.
* B: What is the simplest solution? Making individual a child of aggregate observation?
  + Discussion of organism observation table.
* BP: Are we confusing the data model with the data exchange model?
  + BB: Yes. That’s why we dropped this. We don’t want to impose the data model on the exchange model.
* A: This effects every dataset where you have both aggregate and individual data.
* Aaron: Schedule a call with Nick. Try for Tuesday, 1 PM California is 9 AM New Zealand.
* Aaron, Nick, Bob, Mark, Brad need to be on the call.

**Tree Traversal algorithm**

* Brad thinks Aaron’s approach would work.
  + Brad hasn’t seen a many to many table for this problem. Could be a nice little publication.
  + A: Many to many table is essentially an unsorted array.

**Data Loading**

* BB: How is data loading going?
  + A: Fairly straight forward as long as he can determine what a field means.
  + Takes a long time and is cpu bound.
  + Is currently loading 6 data sources and estimates it will take 4 days for them all to load.
* BB: You’re doing it record by record, carefully.
  + We did that at UA initially.
  + Switched from record by record to big SQL statements that load in bulk.
  + Reduced time from 3 days to a few hours.
* A: We insert in a hierarchical way. So can’t do a bulk insert. Have to insert the tables that other tables are dependent on first.
* BB: Continue as you are doing. We can deal with optimization later. Slow is ok for now.
* A: We’re loading 5 of the 9 sources now.
  + Remib, NCU-NCSC, UNCC, entrish(?), still outstanding.
* BB: Can we get a DwC dump on those data sources that Bob is involved with?
  + BP: Will see.
* BP: Are we still looking for other data sets? Just got an Arctic plot data set.
  + BB: yes.
  + BB: Eventually a Canadian aggregate source is supposed to be available all DwC compliant.
  + A: What’s it called?
  + BB: Canadensys – virtual herbarium for Canada
* BP: Inclined to forget about UNCC and just take what is in UNC.
* BB: Re the data hole in Canada. We could search the gbif data for the acronyms. Brad could tell which are providing data to gbif, and go after those providers that aren’t serving to gbif.
* A: Found how to get DwC dump from Canadensys.
  + BB: That’s great.
  + We need not to duplicate what is being provided to gbif.
  + A: Suggests we just flag the duplicates.
  + BB: One record still has to be tagged as the authoritative record. Thinks Canadensys record should be the authoritative source since it’s closer than gbif to the original record.
  + BB: Aaron, see if there is a difference between the Can and gbif data.
* BB: How long until everything is loaded?
  + A: Physical loading is taking a week. Then he needs to do the mapping (manual). Will take a while to develop the mapping for the other sources.
* BB: Is the current loading to VegBIEN?
  + A: Loading directly to VegBIEN
  + Map to standard DwC and to VegBIEN
  + BB: You are mapping new fields as they come up?
  + A: Yes if don’t yet have a VegBIEN mapping.
* (Discussion of how to store cultivated, exotic; ultimately, no change to VegBIEN schema.)
  + A: Actually, cultured is a growth form field, not an explicit field.
  + BB: Need an “is cultivated” and “cultivated reason” (how the info was discovered)
  + A: Where to add those?
  + BB: It’s an attribute of the occurrence.
  + BP: Agrees it’s needed.
  + BB: Add it to taxon occurrence or \_\_ so can be used by plots or specimens.
  + BP: native, exotic, cultivated
  + BP: growth pertains to the taxon
  + Cultivated, etc pertain to the occurrence.
  + BP: Now sees the logic. OK as is.
  + BB: So the attributes for cultivated are already there.
  + BB: Taxon class: a given taxon occurrence could have one, none or many applied
  + Mike wanted these attached to aggregate occurrence.

**Documentation of VegBIEN schema**

* BP: How will we generate the documentation of the schema?
* A: Convert the data dictionary into comments then, convert to graphical format.
* BB: Much of it can be done automatically, then all of us can work on it to finalize it at end of year.

**Priorities for next week**

* Finish loading DwC data priority.
* Schedule conference call with Nick, hopefully on Tuesday.
* Start implementing Aaron’s tree algorithm.
* If there’s time, summarize the Canadensys data so Brad can prioritize which data providers to go after.
  + A: May not get to that until later.
* BB: Then validation will be next.
* A: Tree algorithm – start implementing?
  + BB: Go ahead. Can always revert to the other algorithm is yours doesn’t work out.

**To Do List**

* Finish loading DwC data priority.
* Schedule conference call with Nick, hopefully on Tuesday.
* Start implementing Aaron’s tree algorithm.
* If there’s time, summarize the Canadensys data so Brad can prioritize which data providers to go after.