12-04-09 BIEN db

Agenda

- How's DwC loading progressing? (database tuning, REMIB, DiGIR client application)

- VegX changes

- VegBIEN validation utilities

DiGIR

Aaron wrote his own DiGIR client in Python

Means we can get DwC data from servers ourselves.

REMIB

A: How many nodes are live and which are static (120 nodes). Many nodes are down.

Many are DiGIR servers.

GBIF data

MS: We’re aggregating, as GBIF is.

Perhaps we should consider this as a view rather than harvest phenomenon.

Everyone has to fix the problems.

BB: Thinks REMIB doesn’t share with GBIF.

MS: Not sure what GBIF considers it’s relationship to BIEN to be?

BB: It was a personal arrangement through Naija.

Should probably talk with Dave Remsen.

MS: What is BIEN’s relationship with museums (eg Mex)

BB: Most botanical collections prefer people to get the data directly from them.

Should contact Dave Remsen.

MS: The museum record domain overlaps with GBIF. The ecological data don’t.

BB: Agrees that it would be good if Aaron could spend more time on plot data.

Having a small set of tools that can ingest specimen data will be very useful.

BB: What specimen sources will be missing from what was in BIEN 2?

Goal is to be data source complete before moving on to plot data.

A: We have all specimen data imported.

Under a week for everything except SpeciesLink

Database tuning

MS: Postgres tuning, people are dubious of a 10 to 50X speed up just from moving from a VM to real server. Identified some tuning they could do on their VM.

Instead of reloading the current database, try the new configuration on new installation and benchmark it.

BB: Wants Aaron to be able to keep working while db is reloading.

A: Needs all of one data source intact to validate.

Can validate on nemoy.

BB: Need some sort of generic validation.

A: Has new data (Canada and REMIB)

BB: Don’t go beyond the sources you’ve already identified.

A: When to start loading plot data?

BB: Next. Move on to loading plot data.

A: CTFS first?

BB: yes

A: Have a dump from CTFS or could try to import VegX.

BB: Fastest would be to get it directly from database.

A: We’d have to redo the mapping.

Stems are now linked to organisms.

Because VegX is so broad, we will have custom mappings for each source.

BB: Let’s try to stick with VegX for now.

Owe it to Shah and Rick to try to import CTFS via VegX.

A: Probably easiest for us to deal with VegX. Need to make sure all the records are in VegX. The files are large and can time out.

BB: May end up wanting DwC archive format (csv).

But we lose the hierarchical schema.

A: CSV is flat and VegX is very hierarchical.

MS: He and Aaron should talk about CSV serialization of VegX.

BB: For now let’s use VegX.

MS: What about TurboVeg?

BB: Need to ask Bob.

MS: Are there a lot of new world plots? In TurboVeg?

BB: No.

MS: Then they’re not important.

BB: Getting new sources is good, but need to make sure we get our existing data in BIEN3.

A: SpeciesLink does record by record import of Brazilian data.

BB: Plots will take a while. Anxious to get validation taken care of.

BB: SpeciesLink is an aggregator and they are very active collecting small Brazilian herbaria data.

BB: Need publications and infrastructure products that let people merge data easily.

More important we do a really good job with fewer sources rather than spending a lot of time going after new sources.

Put the validation utilities on hold until database is assembled.

Get the plot data in first. Do that this week.

Depending on where you stand then, we’ll see about moving on to validation.

Name res and geovalidation are a new layer of information on BIEN db. It’s part of the analytical database, not the original user’s data.

A: Are there revisions to the VegBIEN schema we still want to make?

BB: Keep all tables for now.

BB: National Park data may make use of them. Don’t want to work on that now.

Expect to find the schema weakness as we load data (and break the schema, esp. with plot data).

A: What does breaking the schema mean? Putting things in user defined fields?

BB: Yes or overlooked things that should be first class fields.

MN: What is meant by validation?

BB: First we need to faithfully represent the original data.

Checking for errors of content (TNRS and Geo validation) is part of the analytical database.

MS: Regarding filtered push. Got a response describing “here’s what we’re working on”. But how to push fixes out isn’t done yet.

BB: Those error reports back to the data providers will get us a lot of buy-in.

MS: Need an authoritative gazetteer to compare against. BioGeomancer.

BB: When he and John did geographic name resolution. A lot of the problem was spelling errors.

MS: Isn’t that what BioGeomancer was doing?

BB: Cleaned up geographic names with John, found three types of errors:

Sources of shape files to check point-in-polygon had problems due to variants on polygon names.

Country, state, province are easy.

Checked in the major languages in the Americas: Eng, Span, French, Portuguese

Character coding issues.

Make it python and postgres

There need to be a reliable, stable web services for place names.

A: Database tuning, do what?

SpeciesLink is still running on only one core, so slow. Do we want to parallelize it?

Would use a parallel python package to import data. All data sources could use it.

MN: Ask Mark and Jim those two questions.

BB: Importing CTFS data is highest priority.

Spend 5-10 percent of time on optimization efforts.

Half time on each.

Automating validation of name sources.

A: Where are John’s scripts for geovalidation?

BB: Not sure where, but on Redmine.

A: TurboVeg?

BB: Don’t worry about turboveg (old world data).