**March 9, 2012 BIEN Database Conference Call**

**Participants:** Aaron Marcuse-Kubitza, Mark Schildhauer, Brad Boyle, Martha Narro

**Unavailable:** Jim Regetz, Bob Peet, Mike Lee

**Agenda**

* General Business\*\* Next week the call switches to 1 PM for Tucson folks due to Daylight Savings Time.
* VegX
* Data Loading

**Notes**

**VegX**

* A: When do we want to start changing the XML schema?
  + It will be somewhat time consuming to do it.
* BB: Let’s warp up DwC mappings, get all specimens loaded first.
* MS: We’ve got go ahead to fork VegX.
  + Then approach TDWG and VegX to get our changes reincorporated into the VegX standard.
  + So next steps are:
    - We can proceed unilaterally to propose changes.
    - Open the door to make them part of the standard.
* BB: Yes, Aaron needs to refactor VegX until it works, then ask for Nick and Bob to help make this part to the standard.

**Getting Feedback to Data Providers**

* BB: Sent Shelly McMahon the error reports.
  + Need a newer dump from UA herbarium (SALVIAS) or a way of automating it.
  + Brad will break the ice with Shelly about possibility of automating, then hand it off to Aaron.
  + Thinks the error reports will get us buy in from data providers.
* MS: It could be a good time for us to look into James Macklin’s filtered push.
  + We’re providing feedback to data providers which is part of what filtered push does.
* BB: We’ll need to be careful.
  + We don’t want to report errors related to BIEN’s schema. People don’t need to know about those types of errors.
  + We want to report errors that have to do with their data.
  + Structural issues, they don’t want to know about
  + Content errors, they do, such as corrupt data.
  + Nice if there were a webpage with errors from their last data upload.
* MS: Also could provide a summarization of their data.
  + Even simple things like number of distinct species in their dataset is something people often don’t know.
* BB: Great idea.
  + Also return the errors with identifiers so their technical people can easily identify the problematic records and correct them.
* MS: Aaron’s page on the UA errors is a good start on this.
* BB: Yes, but Aaron is doing it manually. We need to automate it.
* A: Has a tool that does some of it.
* BB: Building a tool that automates reports may be out of scope.
* BB: In interim, where should we put this information?
* A: Currently creating a wiki page for each data provider.
* BB, MS: That’s fine for now.
* BB: Want to automatically send them an email when there’s new information.
* MS: It’s a good time to check in with Macklin and Morris on state of filtered push work. They may have a mechanism for reporting back about when new errors are discovered.
* BB: Are you suggesting they may have a web service for that and we should use it?
* MS: Yes.
  + We should remind them we’re encountering these issues and let them we want to know how Filtered Push is progressing.
* BB: Morris retired and hanging out at Harvard Herbarium. Seems to be busy as ever.
* MS: Will contact them.

**Modeling Political Divisions**

* BB: Where do we stand on how how we model political divisions?
  + Looks like VegBank doesn’t have recursive a model.
* A: Was going to use the recursive model described in the email.
* BB: Has implemented it using right and left indices – allows one to retrieve all parents and children without having to walk up or down the tree. It greatly decreases the select time, but increases the insert and delete time.
* MS: Is there a name for this approach? How to maintain hierarchical relationships in relational databases has always been a challenge.
* BB: It’s called modified tree traversal. Works as long as there are no circular references in the tree.
* MS: Is that what Tree Base uses?
* BB: Don’t know.
* MS: What has iPToL project used?
* Martha: Will ask Naim if iPToL is using something to traverse phylogenetic hierarchies in relational databases.

**Loading Data**

* BB: Aaron, for those data providers serving dwc, try to get permission to access those digger servers.
  + NYBG would be a good place to start. Bet even they won’t give us everything they are providing to GBIF.
  + We’ll have to live with some filters (data limitations).
* BB: As you move through the specimens, you’ll see the smaller data providers may not be using dwc and you’ll need to map those data yourselves.
  + In the old BIEN 2 scripts, data in file loads into mysql table another script loads into a standard staging table. Staging table then loads to database.
* A: We don’t use the staging table. Load database directly.
* A: What priorities do we want the validations to have?
* BB: Load then validate.
  + Import, map, load, validate to check if it went correctly.
  + Build the error report pages as you have been.
* A: Want any changes needed to the way he’s been formatting error reports?
* MS, BB: Nope. They’re fine.
  + The task of building a more robust error reporting should wait until after \_\_\_\_ this summer.
* A: Next step is importing next data sources.
* Yes
* MN: Just loading specimen data now, right?
* BB: Correct.

**To Do**

* Aaron:
* 1. implement place rank hierarchy
  + use GML/DwC standard ranks
  + each record has a parent down to the root
* 2. load remaining DwC inputs
  + recreate verbatim staging DB from flat files
* 3. refactor VegX XML schema
* 4. automated feedback mechanism
  + triggered when an import is run
* Mark: Contact Maclin and Morris to get update on status of Filtered Push
* Martha: See if the iPToL group has used for representing and retrieving heirachical data (phylogenetic trees) from relational databases.