VegBank - Bug #1079

Revise DB model to better deal with idiosyncratic taxa

05/29/2003 05:28 PM - Michael Lee

Status:	Resolved	**Start date:**	05/29/2003
Priority:	Immediate	**Due date:**	
Assignee:	P. Anderson	**% Done:**	0%
Category:	misc	**Estimated time:**	0.00 hour
Target version:	UnMilestoned	**Spent time:**	0.00 hour
Bugzilla-Id:	1079		

Description

We need to allow users to enter "weird" taxa without cluttering up the plant taxonomy module. We can do that by making the taxonObs/taxonInterp look more like commClass & commInterp. Here is the email from Bob describing this. We need first to agree on what the model needs to look like exactly, then implement it. High priority after release 1.0.

From peet@unc.edu Thu May 29 20:06:46 2003
Date: Thu, 29 May 2003 03:26:16 -0400 (EDT)
From: Robert K. Peet <peet@unc.edu>
To: ...
Subject: irregular taxa.

Proposed change in VegBank with respect to handling of irregular taxa.

I have become very worried about cluttering up the plants portion of the database with irregular taxa that are idiosyncratic to particular investigators. Here is what I think we should do.

When an investigator has an irregular taxon such as " Carex-fuzzy" , "Potentilla canadensis + P. simplex", or Astilbe + Aruncus these should be handled in such a way that they do not end up in the plantName and plantConcept table. Each plant should be linked to the lowest level concept in the plantConcept list that is certain. In addition, the idiosyncratic names should be stored in the taxonObservation and a capability of linking the taxon to multiple concepts with a lower certainty should be provided. For example "Potentilla canadensis + P. simplex" would have a high certainty link to Potentilla, and some lower quality links to each of P. simplex and P. canadensis. To realize this we need the following changes.

TaxonObservation.
Change field PlantName_ID to PlantName and change the contents from a FK to text. This had all along been our intent and I don't know when we slipped and made this a foreign key. This has some substantive ramifications for loading data that need to be worked out.

Split TaxonInterpretation into two tables: TaxonClass being a child of TaxonObservations and TaxonInterpretation being a child of TaxonClass. Their contents would be as follows.

TaxonClass
PK
FK-taxonObservation_ID
interpretationDate
FK-reference_ID
FK-party_ID
FK_role_ID
interpretationType
originalInterpretation
currentInterpretation

notes
notesPublic
notesMgt
TaxonInterpretation
PK
FK-taxonClass_ID
FK-plantConcept_ID
classFit [new field, closed list, similar to commInterpretation]
classConfidence [new field, closed list, as in commInterp...]

On top of this we need a business rule that there must be one TaxonInterpretation that has an absolutely correct classFit. To do this means adding some higher level taxa to the plantConcepts and plantNames. All concepts must ultimately filter up to all plants [plus probably a second class for lichens], below which we can have [Charales, Marchantiomorpha (liverworts), Anthocerotophyta (hornworts), Lycopsida, Equisetopsida (horsetails), Filicopsida (ferns), and Spermatopsida. Below Spermatopsida we would have Cycads, Conifers, Ginkgos, Gnetales, and Angiosperms. Below Angiosperms we might have Nymphaeles, Austrobaileyales, Eudicots, Monocotyledons, Chloranthaceae, and Magnoliids.

We could stage the implementation as follows
1. add two new fields to TaxonInterpretation
2. add top to the phylogenetic tree
3. change TaxonObservation:plantName from FK to text
4. split table and support change in data input and table preparation

I am very keen to do steps 1, 2 & 3 as quickly as we can (without holding up release 1). Step 4 can wait until the next release.

Related issues:

Blocked by VegBank - Bug #1169: Upgrade VegBranch to new 1.0.2 Vegbank data m...	**Resolved**	**10/02/2003**
Blocked by VegBank - Bug #870: New XML export and import into the VegBank system	**Resolved**	**11/13/2002**

History

#1 - 05/29/2003 05:30 PM - Michael Lee

commenting on Bob's comments above:
step 1: the 2 fields are fit and confidence? This needs to wait a bit, as it is a model change.

step 2 is a loading issue and not a database schema issue, so it could be addressed quite quickly.
step 3 is actually already in place, meaning that the string is stored in taxonObs as an implementation field that John set up. But it will need to change its name, probably. (now called cheatplantname) Then dropping plantname_id from the table shouldn't be too big of deal.

step 4 is obviously a larger model change and must wait.

Bob added: For info on how to nest families into my angiosperm types, one of us will need to check the tree-of-life website.

#2 - 07/01/2003 11:37 AM - Michael Lee

I like this approach, but I'm not sure that it will adequately solve the problem of P. simplex + P. canadensis. There are (at least) 2 different ways one might want to link to both of these 2 species. Case 1) I think that the species I observe on this plot is all the same species, but I can't determine whether it is P. simplex or P. canadensis. It is one or the other. Case 2) I cannot tell the difference between P. simplex and P. canadensis on my plot, so this concept is an aggregate of the two species, sort of a taxon at a level :section or subgenus.

I don't know if the distinction between the two is really that important to capture. Bob?

One might solve this with a field in the taxonClass table called taxonClassType which would describe how the concept observed on the plot relates to the set of plantConcepts referenced in taxonInterpretation. Possible values would include:
exact [perfect match to one plantConcept],
aggregate [union of multiple plantConcepts referenced],

one-of [it's one (and only one) of multiple plantConcepts],
subset [the ONE plantConcept referenced is larger than what we'd like, but it's the best we can do. in the above example, this would be for the record that links to Potentilla]

This new field (taxonClassType) is similar to classFit, but classFit deals more with the quality of the fit as perceived by the interpreter. taxonClassType is more of a switch to let us know how to handle the possible multiple concepts listed in taxonInterpretation. Perhaps we don't need both of these fields, though, as commInterpretation is a more challenging task: how well does this community fit the "ideal" community, whereas plant taxonomy is (sort of) more concrete.

#3 - 07/04/2003 03:58 PM - Michael Lee

"stage 2" from a comment of Bob's above is completed, that is the top of the phylogenetic tree has been loaded into VegBank. Though there are some differences in the USDA stance and his (ie liverworts=Hepatophyta, not Marchantiomorpha). Someone who knows something about the plant classification hierarchy (ie not me) should look into the top of the hierarchy as loaded on VegBank and make sure that it's ok. Kingdom through Class for Plantae are pasted at the bottom of this email.

excerpts from an email from mtl, July 3, 2003:

I have captured and loaded into vegbank the entire USDA hierarchy as of today's date. They have all of Plantae and the lichens of Fungi available, and this is what I have loaded. This should help us when dealing with idiosyncratic taxa.

Feel free to check it out with the plant query:
http://vegbank.nceas.ucsb.edu/vegbank/forms/plant-query.html

for example:
http://vegbank.nceas.ucsb.edu/vegbank/servlet/DataRequestServlet?requestDataType=plantTaxon&requestDataFormatType=html&clientType=browser&taxonName=Magnoliophyta+&taxonNameType=&taxonLevel=&targetDate=&party=

and
http://vegbank.nceas.ucsb.edu/vegbank/servlet/DataRequestServlet?requestDataType=plantTaxon&requestDataFormatType=html&clientType=browser&taxonName=Cyperales+&taxonNameType=&taxonLevel=&targetDate=&party=

and so on.

-------USDA plantae Kingdom through Class----------
Kingdom Plantae -- Plants
Division Anthocerotophyta -- Hornworts
Subdivision Anthocerotae
Class Anthocerotopsida
Division Bryophyta -- Mosses
Subdivision Musci
Class Andreaeopsida -- Granite mosses
Class Bryopsida -- True mosses
Class Sphagnopsida -- Peat mosses
Division Hepatophyta -- Liverworts
Subdivision Hepaticae
Class Hepatopsida
Subkingdom Tracheobionta -- Vascular plants
Division Equisetophyta -- Horsetails
Class Equisetopsida
Division Lycopodiophyta -- Lycopods
Class Lycopodiopsida
Division Psilophyta -- Whisk-ferns
Class Psilopsida
Division Pteridophyta -- Ferns
Class Filicopsida
Superdivision Spermatophyta -- Seed plants
Division Coniferophyta -- Conifers
Class Pinopsida
Division Cycadophyta -- Cycads
Class Cycadopsida
Division Ginkgophyta -- Ginkgo
Class Ginkgoopsida
Division Gnetophyta -- Mormon tea and other gnetophytes

Class Gnetopsida
Division Magnoliophyta -- Flowering plants
Class Liliopsida -- Monocotyledons
Class Magnoliopsida -- Dicotyledons

#4 - 07/07/2003 04:01 PM - Michael Lee

I will outline 3 different options we have for implementing an ability to handle idiosyncratic taxa, from least database schema changing to most. Let me first say what the goals of the change are:
goal 1) allow taxonObservations to be categorized to different degrees of fit and confidence
goal 2) allow author to describe a plant with any name, but do so without adding "weird" names to plant taxa module
goal 3) make the treating of plant taxa interpretations more parallel to community interpretations
---

--option A) the "mid-80's Honda": change nothing in the database model--------
Use extant field "cheatplantname" that is currently an implementation field to store author's plantName. PlantName_ID is not currently required in the implementation model, so this could be left blank where the name is not standard. This accomplishes goal (2). Goal (1) can be accomplished by adding new value options for the closed list on taxonInterpretation.interpretationType. Current options are: author, computer generated, simplified for comparative analysis, correction,finer resolution.
To this we could add (I'm not sure of the exact values we should use for this, so I use very wordy examples to make the point clear):
--author, less precise, but very certain
--author, more precise, but less certain
The former would be, in the case of P. simplex + P. canadensis, the genus Potentilla. The latter would be used for 2 records, one linking to P. simplex and the other to P. canadensis. So for the single taxonObs, we would have 3 taxonInterpretation records.
For non-weird taxa, we would still use "author" which would indicate that we should assume that the plant is of good precision and good certainty.
This would accomplish goal (1) in that weird taxa can be recorded and linked to taxonObservations in such a way that we can understand what taxa it might be and what taxa it certainty is.

---------END OPTION A----------

----OPTION B: the "family car": add 2 fields to TaxonInterpretation--------
This option is the same as option A except that instead of adding values to the closed list for taxonInterpretation.interpretationType, we would add the 2 fields to this table: classFit and classConfidence. This accomplishes goal(1) better than option A), but at the expense of a model change. It also moves a step closer to goal (3) in that it mirrors commInterpretation more.
The values for classConfidence would be identical to commInterpretation.classConfidence (High, Medium, Low), but we still have the problem that these are not defined particularly well. High would be for "Potentilla" in the above example and Medium(?) for "P. simplex" and "P. canadensis". "Low" would be used for flat out guesses?
classFit is, in my opinion, somewhat different from commInterpretation. The problem lies in the idea that plots can represent communities to a greater and lesser value; it's a gradient. There are some plots that really represent community X extremely well and there are others that are still community X, but to a lesser degree. The fit scale on commInterpretation (Absolutely wrong, Understandable but wrong, Reasonable or acceptable answer, Good answer, Absolutely right) works well with communities.
I don't believe that it works as well with plant taxa. Borders may be grey for communities, but they **attempt** to be more black and white for plants. A plant is either "Red Maple" or it is not (ignoring for the moment hybrids). It can work in that "Pine tree" is absolutely wrong for red maple, "Viburnam acerifolium" is understandable but wrong, "Dicot" is acceptable, "Maple" is a good answer, and "Acer rubrum var. rubrum" is absolutely right. But this is dealing less with fit and more with resolution of the taxon in question, and the boolean statement of correct or not correct. Resolution can be looked up in plantStatus.plantLevel if one wants to, and that leaves the fit field to the boolean statement of correct or not. Can someone come up with some values for classFit that make sense for plant taxa?

---------END OPTION B----------

-----OPTION C: "the Cadillac": make all the changes----------

split taxonInterpretation into 2 tables: taxonInterpretation and taxonClass, which make goal (3) very much realized. Still have the problem with classFit described in option B. Very expensive in terms of database changes, mostly in reworking code that deals with taxonInterpretation table now. Adds the elegent option of clearly lumping taxa or making it a choice of taxa with the field "taxonClassType" in taxonClass. Not possible to add this field without splitting the tables, really, except with some strong business rules. See the .csv attachment that shows the fields to add/change for taxonInterpretation.

---------END OPTION C----------

#7 - 07/12/2003 02:46 PM - Robert Peet

I regret to report that as best I can tell Michael's options A and B in Note #4 fail because they allow multiple records for the observation of a unique taxon in a plot observation. We thus lose track of just how many taxa really occur in the plot observation. We could hack this a bit by adding a new field to indicate the first observation of the real taxon, but this is so highly inelegant that I am sure it would cost us a lot in the long run. I vote for fixing this ONCE and only once with the correct solution C. We can discuss whether this should be immediately or in the next release, but any other solution that I can think of is going to simply waste time and put opff the inevitable.

#8 - 10/02/2003 11:29 AM - Michael Lee

update change XML document to reflect what we will change. Then create "ALTER TABLE" Sql to be run against vegbank. DO NOT RELOAD / REBUILD ENTIRE DB.

#9 - 10/06/2003 10:52 AM - Michael Lee

We suggested making a recursive key in TaxonInt to allow a link to "otherTaxonInt_ID" which could extend an interpretation and allow one to interpret as P. simplex then P. canadensis as part of the same TaxonInterpretation. however, we could like the user to also interpret as "Potentilla" which would have to be a second taxonInt record. Inelegant. I dislike recursive keys, too.

Better would be a new table called taxonGroup which would link back to TaxonInt and to PlantConcept, allowing any number of taxa to be mentioned as part of a TaxonInt. There would still be a link to PlantConcept from TaxonInt so that the "best single taxon" could be linked as part of the same irregular taxonInterpretation. In the above example, "Potentilla" would be linked to directly from TaxonINt, and P. can. and P. simpl. would link via TaxonGroup.

See http://tekka.nceas.ucsb.edu/~lee/newIrregTaxa.jpg for a brief exaplanation and ERD of this.

#10 - 10/13/2003 04:06 PM - Michael Lee

model has been updated. SQL file needs to be run against vegbank to update the structure. As part of QA of XML import, loading taxa that are irregular should be checked. Also, querying of plots should show options for showing irregular taxa in "summary" (only taxonInt.plantConcept_ID) or "details" (all taxonAlt.plantConcept_ID) views.

Passing this bug to Mark, who will run the SQL file in cvs against the database to updated model. Then this bug can be closed.

sql file is in vegbank/veg_plot/sql/vegbank-changes-1.0.1.sql (revision 1.2 committed 10/13)

#11 - 10/20/2003 04:22 PM - P. Anderson

The SQL update file has been run on all 3 machines, gyro, tekka and vegbank. Note that the file no longer lives at vegbank/veg_plot/sql/ and is currently enjoying its new home at vegbank/src/sql/vegbank-changes-1.0.1.sql.

#12 - 01/31/2005 02:53 PM - Michael Lee

changed from components that are to be deleted to "misc" so that bugs don't get deleted with component. Sorry for all the email.

#13 - 03/27/2013 02:16 PM - Redmine Admin

Original Bugzilla ID was 1079

Files

stem_irregTaxa_chngs_3OPTIONS.htm	56.3 KB	07/07/2003	Michael Lee
stem_irregTaxa_chngs_3OPTIONS_fix.html	10.7 KB	07/07/2003	Michael Lee