BIEN 3 - Task #314

Import CTFS data

01/05/2012 03:09 PM - Aaron Marcuse-Kubitza

Status:	Resolved	**Start date:**	01/05/2012
Priority:	Normal	**Due date:**	
Assignee:	Aaron Marcuse-Kubitza	**% Done:**	100%
Category:		**Estimated time:**	0.00 hour
Target version:			
Activity type:			

Description

- coordinate with Shash
- CTFS has a lot of stems data

History

#1 - 01/05/2012 03:09 PM - Aaron Marcuse-Kubitza

- Assignee set to Aaron Marcuse-Kubitza

#2 - 01/09/2012 11:40 AM - Aaron Marcuse-Kubitza

- % Done changed from 0 to 30

I uploaded Shash's VegX file to vegbiendev at /home/bien/svn/inputs/CTFS/src.VegX.xml .

#3 - 01/10/2012 01:20 PM - Aaron Marcuse-Kubitza

Looking at Shash's VegX file, I noticed that she uses different VegX elements than we do for SALVIAS and NYBG data. For example, she stores each organism in an aggregateOrganismObservation, while we use individualOrganismObservation and individualOrganism.

In general, it is possible that different organizations will use different VegX elements for the same data. Thus, we may need to have a separate VegX->VegBIEN mapping for each data source, or possibly a custom-VegX->standard-VegX mapping to convert the data source to our use of VegX.

In light of this, would we prefer to have organizations generate VegX themselves, and then convert their VegX, or just provide their raw data and suggested mappings to VegX? For data verification, it's certainly easier to run queries on a database than on an XML file.

#4 - 01/10/2012 02:20 PM - Aaron Marcuse-Kubitza

e-mail from Mark Schildhauer on 2012-1-10:

I think tt would be best (and certainly most scaleable) if the elements were clearly enough defined so that groups could create their own mappings, rather than our having to do so. If we went the latter route, we'd have to understand the nuances of any additional data we wanted to incorporate into VegBIEN, rather than let the experts/stewards make those mappings themselves as appropriate. It strikes me that errors and inconsistencies in mapping will arise primarily due to ambiguity in the definitions of the fields, and lack of adequate examples. Especially, e.g. when we are looking at inconsistent useage of high level terms like "aggregate" vs "individual"!! I was similarly concerned about having solid definitions of the terms in the new table structures you are actively developing based on the whiteboard revisions from the last meeting...

#5 - 01/11/2012 11:08 AM - Aaron Marcuse-Kubitza

e-mail from Brad Boyle on 2012-1-10:

Interesting observation. I must admit, I haven't inspected Shash's VegX yet, but given your comments I will definitely do so asap.

I was expecting Shash (and Steve/Rick) to provide us with a new version of the CTFS data, with observations at the individual level. Something you may not know is that the CTFS data originally loaded to BIEN2 is aggregate data. Rick (via Steve and his students) did not provide us with the original data; only an aggregated summary. I was expecting Shash to load the raw, individual-level data to VegX, but it sounds like she re-loaded the old aggregated data. (This would also explain Steve's comment about providing you with a flat file of the raw data, rather than a connection to the original data; it would be very difficult to represent the raw CTFS data as a flat file). Without having inspected it myself, I suspect Shash's use of aggregate observations may not be incorrect. Although I am disappointed that Rick did not provide individual-level observations. Anyway, I will have a look before I say anything more.

#6 - 01/12/2012 04:37 PM - Aaron Marcuse-Kubitza

e-mails on 2012-1-12:

Steve Dolins:

We used the identical aggregate data file used for BIEN I... I think the data file was generated by Rick on November 12 2009.

I don't think it will be a significant amount of work for Shash to generate a VegX file with raw data. She has completed the mapping and I believe she has written a program to generate the file from a dump of the view table, DFTemp. Again, I don't think we should do more than one plot... Let Nick and Rick validate.

Rick Condit:

I thought we had already covered all these points. I was hoping Shash
would write a script to map DFtemp, every individual record, to vegX.
Best if it's all the plots there, not just one plot.

Brad Boyle:

Thanks Rick, that was what I was expecting too. I would be totally on board with having all the plots so we can do a really robust test of BIEN3, and have all the metadata on hand so we don't have to go bugging you for it later.

#7 - 01/19/2012 11:30 AM - Aaron Marcuse-Kubitza

- % Done changed from 30 to 50

I have data from Shash and am working with her to make sure we can parse it

#8 - 09/17/2012 07:03 AM - Aaron Marcuse-Kubitza

- Status changed from New to Resolved

- % Done changed from 50 to 100