BIEN DB Meeting Notes – Brad

Thursday, January 12, 2012

Mapping tool

* A "nice to have", and would greatly aid Aaron in mapping new (plot) data sources
* But would be a big time sink (Mark; agreed by all)
* Out of scope for this year; might be good to seek separate funding to develop this tool
* Caveat: unless Aaron can identify some relatively "ready to use" tool to assist mapping (CLIO? modify VegBranch?)

Web interface

* Postpone for now decision about platform, re-use of VegBank interface, who does it and when. After we have finalized database schema will discuss again.

Mapping and import procedure

* Specimens:
	1. sourceDb🡪DwC\_extract🡪(VegX)🡪VegBIEN
	2. sourceDb🡪DwC\_extract🡪VegBIEN
	3. sourceDb🡪DwC\_extract🡪DwC\_BIEN🡪VegBIEN
	4. sourceDb🡪DwC\_extract🡪DwC\_BIEN🡪VegX🡪VegBIEN
	+ Currently Aaron does #1, using VegX as path only (doesn't actually convert to XML
	+ Should consider using different approach, either #3 or #4
	+ Advantages of this approach is that the only arrow that would need to be customized for different data source is from DwC\_extract🡪DwC\_BIEN. Because DwC is simple, this task would be simple.
	+ DwC\_extract is provided by data provider
	+ All remaining arrows (DwC\_BIEN🡪VegX🡪VegBIEN) would need to be done only ONCE and never changed again.
	+ Therefore we need to specify a the particular version of DwC to be used by BIEN
	+ Brad will do this, and get back to Aaron
* Plots

1. sourceDb🡪VegX🡪VegBIEN

2. plotDB🡪plot\_extract🡪VegX🡪VegBIEN

* + The arrow from VegX🡪VegBIEN should be completed asap, and should be comprehensive, so can be used for all data sources without modification. This will require completion of : (a) revision of VegX schema (especially, to accommodate stems within individuals) and (b) VegBIEN (nearly complete, as per Bob's suggestions)
	+ The arrow TO VegX is proving problematic because of the greater complexity and variability of plot data, and because of the reluctance of many data providers to provide access to raw data (path #1) and instead provided aggregated extract (path#2). We need to handle both, but prefer path#1.
	+ In practice, Aaron has been taking both approaches, depending on what is provided by data provider. Path #1 is preferable, however, as structure of data is represented within database schema. Path #2 requires some guesswork and a lot of emails to determine relationship between elements in plot\_extract, even when Aaron has access to schema for source Db. We should make effort to obtain access to original database whenever possible.

Summary of priorities, in order (Brad's view):

1. Complete restructuring of VegBIEN according to Bob's suggestions, including converting user-defined fields to first class fields, as needed
2. Restructure VegX to support individuals and stems (essentially, same changes as made to VegBIEN). Convert user-defined fields to first class fields, as needed
3. Produce single comprehensive script for importing all elements in (revised) VegX into (revised) VegBIEN
4. Settle upon version of DwC to be used by BIEN ("DwC\_bien"). Brad will prepare revised DwC schema, and Aaron will revise specimen import scripts to allow for new path of mapping as follows:

sourceDb🡪DwC\_extract🡪DwC\_BIEN🡪[VegX🡪]VegBIEN

We still need to decide if mapping is via VegX, or if we go directly from DwC\_bien to VegBIEN. If the latter, then we would have two (and only two) routes into VegBIEN; one via VegX (for plots) and another via DwC\_bien (for specimens)

1. Reimport all data using revised schemas, in this order:
* SALVIAS
	+ Revise mapping to revised VegX
	+ all elements, including stems
	+ complete validations with help from Brad
* NYBG
	+ Revise mapping to DwC\_bien
	+ Import all
	+ Complete validations, with help from Brad
* CTFS
	+ Obtain direct access to source Db (Brad will negotiate this)
	+ Complete new mapping to (revised) VegX
	+ Import all
* Complete validations, with help from Rick
1. Meanwhile:
* Brad to provide additional herbarium datasets (DwC extracts)
* Brad to negotiate access to Rainfor DB
* Mark to provide template MOU to assist with obtaining access to data sources