**BIEN database redesign: progress since the 2011 working group**

**23 August 2012**

**SUMMARY**

Under the direction of Brad Boyle, Mark Schildhauer, Jim Regetz and Bob Peet, with coordination from Martha Narro, Aaron Marcuse-Kubitza has been developing a comprehensive vegetation data repository, BIEN 3 (or [VegBIEN](https://projects.nceas.ucsb.edu/nceas/projects/bien/wiki/VegBIEN_schema)), capable of handling all key data elements and structures of plot and specimen data. The database schema is based primarily on [VegBank](http://vegbank.org/vegbank/views/dba_tabledescription_index.jsp?view=index&entity=dba_tabledescription), with significant modifications and extensions. Data is transferred to VegBIEN from specimen databases via the existing Darwin Core (DwC) exchange schema; plot data is transferred using [VegCSV](https://projects.nceas.ucsb.edu/nceas/projects/bien/wiki/VegCSV), a new plain text schema developed by Aaron based on the VegX vegetation exchange schema. VegCSV overcomes several limitations of XML, including heavy memory use, inflexible hierarchies, and cumbersome Xpaths. Currently all specimen data and most plot data have been loaded to the core VegBIEN database; mapping from the remaining plot data sources to VegCSV is nearing completion. Once these mappings are complete, individual data sources (or the entire database) can be loaded and re-loaded rapidly, allowing for ongoing updates as source data is modified. The current complete loading time of 10 hours—which represents a speed improvement of more than 20x over first-generation loading scripts—is the result of extensive testing and optimization resulting in the switch to column-based data normalization using native SQL functions. Error logs are generated during the loading process and can be used to provide feedback to data providers, allowing them to correct errors at the source.

The remaining steps in the BIEN 3 database redesign are (1) developing the analytical database, (2) implementing and automating previously-developed taxonomic and geographic validation scripts, and (3) providing access to the BEIN database via the BIEN website. Target completion time of all BIEN 3 redesign [milestones](https://projects.nceas.ucsb.edu/nceas/projects/bien/wiki/Milestones) is late November, in time for the 2012 BIEN meeting.

**BIEN 3 DEVELOPMENT HIGHLIGHTS**

Below we list highlights of the BIEN 3 database development process. For a detailed account, see the [database development wiki page](https://projects.nceas.ucsb.edu/nceas/projects/bien/wiki/Database_development#Progress-since-the-2011-working-group).

**Current state of the VegBIEN DB**

* The data repository is implemented in a [PostgreSQL](http://www.postgresql.org/) relational database and managed by a layer of [python](http://www.python.org/) services
* The database is hosted at [vegbiendev.nceas.ucsb.edu](http://vegbiendev.nceas.ucsb.edu/) and can be accessed using the instructions on the wiki under [PhpPgAdmin](https://projects.nceas.ucsb.edu/nceas/projects/bien/wiki/PhpPgAdmin)
  + If you have an account on nimoy or vegbankdev, you will also have an account on vegbiendev with the same login
  + BIEN 3-related files are available on vegbiendev in /home/bien/svn
* We currently have 14 datasources in the VegBIEN database: ARIZ, ACAD, CVS, GBIF, MO, MT, NCU-NCSC, NY, QMOR, REMIB, SALVIAS, SpeciesLink, U, UNCC. This includes all BIEN2 specimen, and most plot data.

[**VegBIEN schema**](https://projects.nceas.ucsb.edu/nceas/projects/bien/wiki/VegBIEN_schema)

* The [VegBIEN schema](https://projects.nceas.ucsb.edu/nceas/projects/bien/wiki/VegBIEN_schema) (Fig. 1) has been created by refactoring the [VegBank schema](http://vegbank.org/vegbank/views/dba_tabledescription_index.jsp?view=index&entity=dba_tabledescription)
* The VegBIEN schema supports several new concepts:
  + methods: plot methodology, line-intercept measurements, and size classes
  + taxon class inclusions/exclusions: growth forms and plant concepts sampled/not sampled
  + location determinations: successive remeasurements of plot GPS coordinates, georeferencing info
* The plantname and namedplace tables have been redesigned as trees, with each element pointing to its parent element
  + We are using [closure tables](http://karwin.blogspot.com/2010/03/rendering-trees-with-closure-tables.html), which store all the paths between tree nodes, to speed up queries
  + Our closure table algorithm is in svn under [schemas/tree\_cross-links.sql](https://projects.nceas.ucsb.edu/nceas/projects/bien/repository/entry/schemas/tree_cross-links.sql)
* MySQL Workbench enables us to regularly synchronize the ERD with the SQL DDL (after it's translated into MySQL)

[**VegCSV**](https://projects.nceas.ucsb.edu/nceas/projects/bien/wiki/VegCSV)

* For plots data, we are using a CSV format similar to Darwin Core (DwC) in order to avoid the [difficulties of working with XML documents](https://projects.nceas.ucsb.edu/nceas/projects/bien/wiki/VegCSV_vs_VegX) and the associated long XPaths
* This new format, named [VegCSV](https://projects.nceas.ucsb.edu/nceas/projects/bien/wiki/VegCSV), uses a vocabulary of terms ([VegCore](https://projects.nceas.ucsb.edu/nceas/projects/bien/wiki/VegCSV#VegCore)) from DwC, VegX, VegBank, SALVIAS, and VegBIEN as CSV column names
* It provides a "grab bag" of terms that are easily mapped to, in the same way that specimen data is easily mapped to DwC
* Hierarchical relationships are represented by storing different terms in different tables (csv files) and defining their import sequence

[**VegCore**](https://projects.nceas.ucsb.edu/nceas/projects/bien/wiki/VegCSV#VegCore)**->VegBIEN mapping**

* We have a [mapping](https://projects.nceas.ucsb.edu/nceas/projects/bien/repository/raw/mappings/for_review/VegCore-VegBIEN.csv) from [VegCore](https://projects.nceas.ucsb.edu/nceas/projects/bien/wiki/VegCSV#VegCore)->VegBIEN
* This human-readable version is automatically generated from the [machine-readable version](https://projects.nceas.ucsb.edu/nceas/projects/bien/repository/raw/mappings/VegCore-VegBIEN.csv)
* The import uses the following algorithm:
  1. Generate an in-memory XML template from the mappings
  2. Insert this tree into the database in dependency order (leaves first) using the VegBank XML import algorithm and [column-based import](https://projects.nceas.ucsb.edu/nceas/projects/bien/wiki/Column-based_import)
* To test the import process, login to vegbiendev and run: make test --directory=/home/bien/svn/

[**Column-based import**](https://projects.nceas.ucsb.edu/nceas/projects/bien/wiki/Column-based_import)

* We now import data by column instead of by row, providing altogether a 20x+ speed improvement and taking only ~10 hours, rather than days
* The algorithm also handles many errors server-side using [wrapper functions](https://projects.nceas.ucsb.edu/nceas/projects/bien/wiki/Column-based_import#Special-case-SQL-function), which avoids the overhead of returning to the client for each error
* Details are on the wiki under [Column-based import](https://projects.nceas.ucsb.edu/nceas/projects/bien/wiki/Column-based_import)

**Data provider feedback**

* We are using a new **error-logging and data provider feedback mechanism**, which logs each invalid value instead of each invalid row
  + This eliminates duplication in the logged errors, making it much easier to see individual problems affecting the data
* Invalid values and their corresponding error messages are placed in an [errors table](http://vegbiendev.nceas.ucsb.edu/phppgadmin/display.php?server=localhost%3A5432%3Aallow&database=vegbien&schema=ACAD&table=specimens.errors&subject=table&return_url=tblproperties.php%3Fserver%3Dlocalhost%253A5432%253Aallow%26amp%3Bdatabase%3Dvegbien%26amp%3Bschema%3DACAD%26amp%3Btable%3Dspecimens.errors&return_desc=Back&table=specimens.errors&sortkey=&sortdir=&strings=expanded&page=1), which is an auxiliary table for each datasource where the errors are deposited
* SQL function calls and type casts are wrapped in an exception handler that saves errors into the errors table
* Feedback is now provided on most if not all places where input data causes errors

**Timeline**

* We created [**milestones**](https://projects.nceas.ucsb.edu/nceas/projects/bien/wiki/Milestones) and a [**development timeline**](https://projects.nceas.ucsb.edu/nceas/projects/bien/repository/entry/to_do/timeline.doc)
* We are using [**Redmine**](https://projects.nceas.ucsb.edu/nceas/projects/bien/issues?c%5B%5D=status&c%5B%5D=subject&c%5B%5D=assigned_to&c%5B%5D=updated_on&c%5B%5D=done_ratio&f%5B%5D=status_id&f%5B%5D=&group_by=&op%5Bstatus_id%5D=*&set_filter=1&sort=id%3Adesc%2Cdone_ratio) to track progress on BIEN 3 tasks
* You can watch an issue by clicking the issue, and in the upper-right-hand corner of the content area, clicking Watch

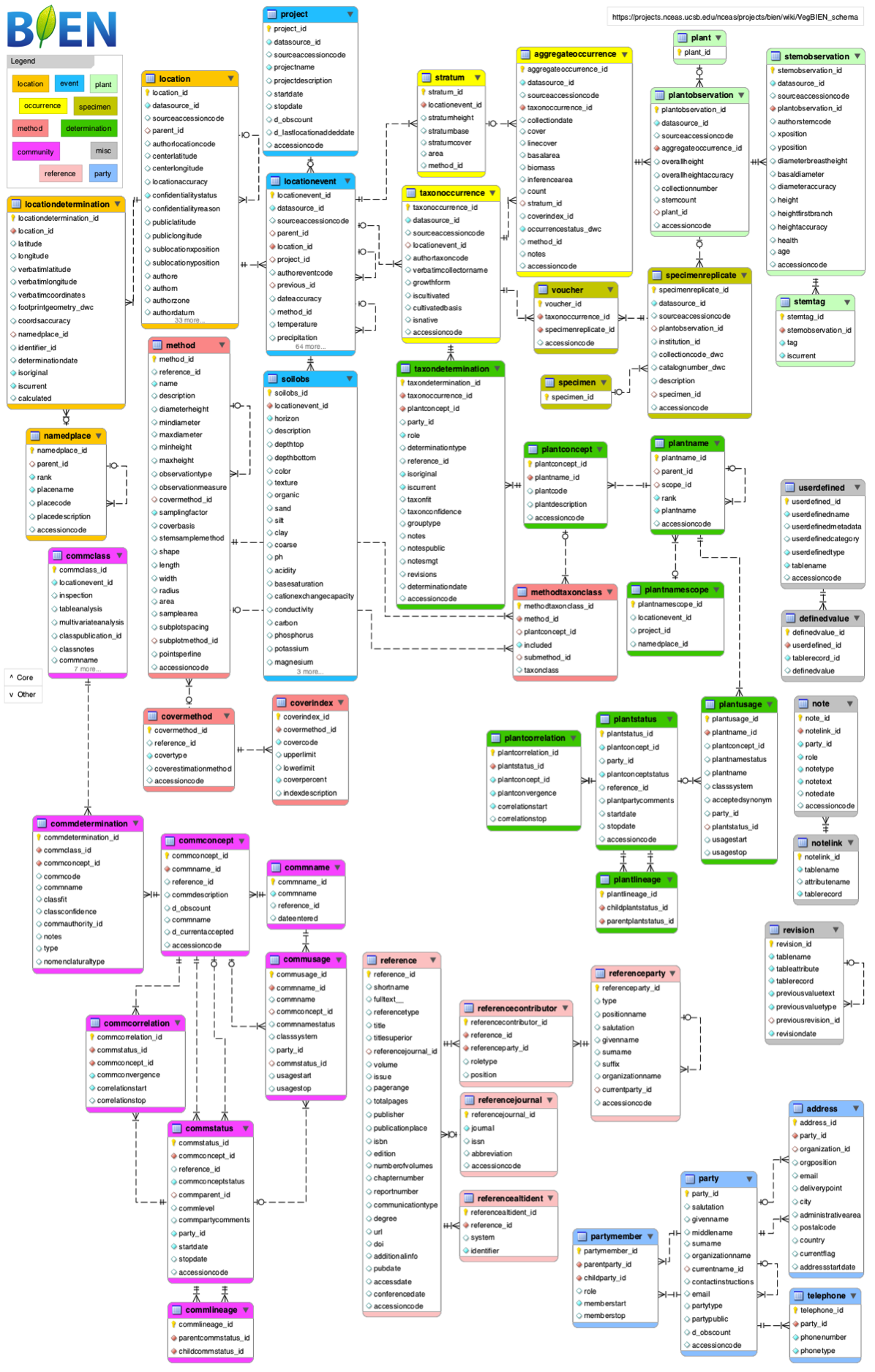


Fig. 1. Current VegBIEN database schema. To examine the schema in detail, go to <https://projects.nceas.ucsb.edu/nceas/projects/bien/wiki/VegBIEN_schema>