**BIEN 3 derived data – Normalized database only**

**Prepared by: Brad Boyle**

**Date: 3 July 2013**

**CONTENTS**

Introduction 2

Overview of BIEN 3 derived data 2

1. Taxonomy 2

2. Phylogeny 3

2.a. NCBI phylogeny 3

2.b. BIEN species-level phylogeny 3

2.c. Detection of non-plant names 4

3. Geography 4

3.a. Geographic name resolution 4

3.b. Geovalidation 4

3.c. New World localities 5

4. Cultivated specimens and anthropogenic (planted, cultivated or otherwise strongly human-influenced) plant community samples 5

4.a. Cultivated specimens 6

4.b. Cultivated, planted and otherwise disturbed plots 7

5. Data provenance, ownership, attribution and access conditions 8

# Introduction

This document summarizes the major high-level standardizations, validations and other derived data products that will form part of the normalized BIEN3 database. These data products are most efficiently calculated either (1) prior to normalization, during the ingest of each data source, or (2) after normalization, as additional normalized values added to the confederated, normalized database.

By "derived", I mean data in addition to the original raw data. These are data based on or calculated from the raw data, but in most cases, not present in the original dataset. By "high-level", I mean that generating these data requires specialized domain knowledge, involves complex algorithms, or both. These attributes of domain-specific knowledge and complexity distinguish high-level derived data from low-level, generic validations and standardizations, such as determining if coordinates fall within the range of valid latitude and longitude (a low-level validation) or converting valid coordinates from degrees, minutes, seconds to decimal degrees (a low-level standardization). As illustrated by the preceding example, "validations" check if the data meet a set rules and flag them accordingly; "standardizations" change the data. Other derived data products include complex data such as species range models; although based in part on the raw data, theuy represent entirely new datasets. Typically complex derived data involves aggregation, such that the derived dataset contains many fewer rows than the raw data.

Not included in this document are derived values more appropriately calculated "on the fly" during construction of one or more materialized views. The calculation of such values is out of scope for construction of the normalized BIEN3 database.

# Overview of BIEN 3 derived data

## 1. Taxonomy

##### Taxonomic names must be corrected and standardized before species observations can be used. Correction include fixing spelling errors and removing extraneous information such as comments and question marks that prevent names from matching. Standardizations include using a only one abbreviated form for an authority, and only one of several variant spellings, even when such alternative are all technically considered correct under the rules of nomenclature.

These corrections and standardizations are performed using the Taxonomic Name Resolution Service (TNRS, <http://tnrs.iplantcollaborative.org/>), a tool developed under a BIEN-iPlant collaboration. Values that should be stored in the BIEN database include:

* The verbatim original name
* The matched name
* TNRS overall match score for the matched name
* Taxonomic status, if provided, of the matched name (i.e., "Accepted", "Synonym", "Invalid", etc.)
* The accepted name; if no accepted name is provided, use the matched name
* Unmatched unmatched terms
* The most probably morphospecies, formed by concatenating the accepted name with unmatched terms, if any.
* TNRS metadata: taxonomic source(s), date of access of the TNRS, versions of TNRS application and database

For the matched and accepted name, in addition to the lowest-level taxon (e.g., species, subspecies, variety, etc.), the following taxonomic levels should also be stored: Family, Genus, Species, Authority of the lowest level taxon.

*Scope of implementation*: Universal. These operations should be performed on the entire database, once all data have been loaded.

## 2. Phylogeny

### 2.a. NCBI phylogeny

To aid in filtering by higher taxon, the final accepted name returned by the TNRS is joined to the most recent NCBI plant phylogeny (Viridiplantae) at the genus level, or, failing a match at the genus level, by family. Taxa in BIEN joining neither by genus nor family are assumed not to belong to the Viridiplantae.

This filtering requires that the Viridiplantae subset of the NCBI phylogeny be cached locally as part of the BIEN database, and indexed in a manner that allowed retrieval of both ancestor and descendent taxa to arbitrary depth. Right and left indexes generated using a modified-preorder tree traversal algorithm are one indexing methods; others are possible.

As an additional shortcut to assist users in filtering by higher taxa, the link to the NCBI backbone should also be used to populate on or more columns with values conforming to a closed vocabulary representing major clades of plants likely to be of interest to researchers. An example of one such division into major clades is "mosses, liverworts, hepatics", "lycophytes", "ferns", "gymnosperms", "angiosperms".

*Scope of implementation*: Universal. These operations should be performed on the entire database, once all data have been loaded.

### 2.b. BIEN species-level phylogeny

Whereas the NCBI phylogeny merely splices BIEN taxa by genus or family onto the NCBI phylogenetic backbone, the BIEN species-level phylogeny is an attempt to resolve relationships among all species in BIEN using a variety of complex algorithms. The BIEN species-level phylogeny is the product of a BIEN-iPlant collaboration, with primary development by Bill Piel.

*Scope of implementation*: Universal. These operations should be performed on the entire database, once all data have been loaded.

### 2.c. Detection of non-plant names

These scripts populate the column `isPlant` by examining the results of the TNRS name resolution output. Match scores for the name itself, the name author, and the family (if included), and reference lists of cross-code homonyms are used to flag the name as definitely representing a plant, definitely not representing a plant, or ambiguous (might be either a plant or an animal, etc.).

## 3. Geography

For BIEN 3, the principal derived geographic data are: geographic name resolution and geovalidation.

### 3.a. Geographic name resolution

##### Geographic name resolution discovers and standardizes names of political divisions within 3-level hierarchy (e.g., Country, State, County). Names are standardized and indexed according to the GeoNames database (<http://www.geonames.org/>). This standardization if performed for **both** BIEN data and the political division names within the GADM political division shapefile database used for the later step of geovalidation (see 3b, below). The end product of geographic name resolution is a standadized names and identifier for each of the three political division Country, StateProvince and CountyParish. These values are required for the subsequent step of geovalidation, which determines if the coordinates of a given taxon observation fall within its declared political divisions. The details of geographic name resolution are provided in Appendix @gnrs.

*Scope of implementation*: Universal. These operations should be performed on the entire database, once all data have been loaded.

### 3.b. Geovalidation

Geovalidation verifies that the geocoordinates (latitude and longitue) of a taxon observation falls within the known limits of its declared political division. Geovalidation is performed as a GIS operation by joining the spatial point (as indicated by its geocoordinates) to the polygon (shapefile) of its declared political division and determining if the point falls inside or outside of that polygon, and, if outside, its distance from the nearest point along the boundary of the polygon.

The political division polygons used for BIEN geovalidation are from the GADM database (<http://www.gadm.org/>). Because the GADM database uses non-standard political division names, GADM names must also be standardized to UniversalNames standards by geographic names resolution, prior to geovalidation (see 3a, above). Geographic name resolution and geovalidation are thus intimately linked as steps 1 and 2 of the same geographic standardization pipeline.

The end product of the geographic name resolution plus geovalidation is two new fields for each of the three political divisions: an integer field indicating if a standard name was matched for that political division, and if so, if the associated point fell in or out of the political division, and, for points falling outside, a decimal value indicating km from nearly boundary point. In addition, a seventh integer field indicates whether or not the record as a whole passed validation. Passing validation requires that the point either fall within all declared political divisions or outside but within an allowable distance from the boundary. Allowing this small amount of error recognizes the inherent imprecision in political shapeful boundaries, thereby avoiding the exclusion of potentially important collections or observations of species near country borders.

Geographic name resolution and geovalidation were initially developed by Brad Boyle and John Donoghue for internal use within the BIEN 2 database, and later improved and optimized by Jim Regetz for BIEN 3. A longer term "nice-to-have" for BIEN will be to provide access to both of these utilities via a publcly-accessible web service and web interface.

*Scope of implementation*: Universal. These operations should be performed on the entire database, once all data have been loaded.

### 3.c. New World localities

Most analysis based on the BIEN 3 database will require filtering out of taxon observations from Old World localities. There, each taxon observation will be flagged is inside or outside the New World, based on a reference list of New World countries. The result is a field called `isNewWorld`, which has a value of 1 for New World localities and 0 for non-New World localities. For BIEN purposes, Greenland is considered part of the New World.

*Scope of implementation*: Universal. These operations should be performed on the entire database, once all data have been loaded.

## 4. Cultivated specimens and anthropogenic (planted, cultivated or otherwise strongly human-influenced) plant community samples

Taxon observations representing plants sewn and cultivated by humans, especially when outside their native range, are a serious source of error for most analyses. For this reason, every effort must be made to detect and remove or flag such observations. To this end, we perform two main sets of validations, one set for specimen observations, and a second set of (currently customized specifically for the FIA database, which is known to contain numerous plots in plantations or forests subject o logging or sivicultural treatments.

### 4.a. Cultivated specimens

Currently, we use a variety of methods to detect and flag specimen observations known or suspected to represent cultivated plants. The end product of this validation is the field `isCultivated` which takes on three values: 0='Definitely knot cultivated", 1="Cultivated", NULL="unknown". These methods were originally developed for the BIEN 2 database by Brad Boyle, and have been transferred mostly unchanged to BIEN 3. The methods are as follows:

**4.a.i. Indications of cultivated status by original data provider.** Unlike most othe BIEN 3 derived data operations, this method must be implemented during the loading stage on a source-specific basis. Different herbaria use different fields or combinations of fields to indicate cultivated specimens, and in fact most herbaria do not record this information at all. Nonetheless, when present, such information should always be preserved as it is the highest-confidence means of detecting cultivated specimens.

*Scope of implementation*: Source-specific. These operations must be performed at the loading stage using rules specific to individual data providers. Applies to specimen observations only.

**4.a.ii. Key words in locality description.** This method searches on the both the locality description and specimen description fields for key words matching a black list of terms possibly indicating that the specimen was collected from a farm, garden or plantation (e.g., "garden", "plantation", etc.). Although this method has a high rate of both false positives and false negatives, it has still proven effective at removing cultivated specimens not detected by other methods.

*Scope of implementation*: Universal. These operations should be performed on the entire database, once all data have been loaded. Applies to specimen observations only.

**4.a.iii. Proximity to herbarium.** This method was designed specifically to remedy the problem of a large number of vouchers specimens of (cultivated) plants from botanical gardens. Lacking a database of the locations of botanical gardens, we instead used locality information for world herbaria from Index Herbariorum, removing from the database (or rather, flagging as cultivated) any specimen observation within 3 km of a herbarium. As many botanical gardens are associated with herbaria, this approach has proven effective at removing cultivated specimens not detected by other methods.

*Scope of implementation*: Universal. These operations should be performed on the entire database, once all data have been loaded. Applies to specimen observations only.

**4.a.iv. Flagging of species outside their known range.** A "black list" of countries where a given species is known not to occur is used to mark species as cultivated. For example, pines do not occur naturally south of Nicaragua; therefore, any occurrence of a pine in a country south of Nicaragua is flagged as cultivated.

Currently, this method has been implemented on a very limited scale for a small set of higher taxa, such as pines and oaks, with well-known country-level distributions. In the future, it may be worth using a more detailed "white list" approach, in the form of state or county distribution records. Although this approach would provide much more thorough filtering of cultivated specimens, the danger of false positives should not be ignored. The user could have the option of inspecting and, if desired, retaining observations which he or she believes to be wrongly labelled as cultivated. Such record-by-record inspection may not be feasible for large data extracts, and the extremely damaging effects of even a single cultivated specimen should be weighed against the slght loss of information due to rejecting a few records from a large dataset.

*Scope of implementation*: Universal. These operations should be performed on the entire database, once all data have been loaded. Applies to specimen observations only.

**4.a.v. Other methods for detecting cultivated specimens.** One other potential method for excluding cultivated specimens, not yet implemented, is the use of outlier analysis to detect and flag unusual occurrence records. However, given the many tropical species that are know only from a handful of widely dispersed occurrences, the consequences of false positives are not trivial. This approach is perhaps be best implement for species with collection records above a certain minimum sample size threshold.

### 4.b. Cultivated, planted and otherwise disturbed plots

Currently, this operation applies only to plot observations from the Forest Inventory and Analysis (FIA) database. Numerous plots in this database—perhaps the majority—represent plantations, logging operations or silvicultural treatments and therefore must be excluded for most analyses. Originally implemented in BIEN 2 by Brad Boyle, with assistance from Bob Peet and Brian McGill, this algorithm with revised extensively for BIEN 3 with assistance from Jes Coyle, Bob Peet and Margaret Evans.

This operation uses a complex series of rules based on metadata specific to the FIA database, in particular the COND table, and is therefore implemented at the loading stage for this source only (see Appendix @FIA for details). The rules are applied to entire plots, not to individual taxon observations. In BIEN 3, the offending plots are marked isCultivated=1, as for cultivated specimens. The majority of plots in FIA are quality as "cultivated" for one or more reasons, and user should take great care to exclude these samples.

*Scope of implementation*: Source-specific. These operations must be performed at the loading stage using rules specific to individual data providers. Applies to specimen observations only.

## 5. Data provenance, ownership, attribution and access conditions

##### Information on data provenance, ownership, attribution and access conditions is central to BIEN's goal of encouraging collaboration while respecting intellectual property and data access conditions. Furthermore, proper attribution is a requirement of access to data provided by GBIF, the major source of specimen data used by BIEN. For this reason, the BIEN schema must support tracking of following information:

* Immediate data provider: who gave the data to BIEN?
* Immediate data provider type. Are they an indexer (aggregator) such as GBIF, or a primary data provider such as Missouri Botanical Garden?
* If the data provider is an indexers, who are the primary data providers of each component dataset?
* Intellectual property. Who are the data owners? In some cases, this may be the same as the primary data providers, or there may be multiple data owners in addition to the primary data provider.
* Are any data access conditions attached by the primary data provider? Other data owners? Data indexers?

Although we may lack some or most of this information for a given data set, we should have somewhere to put it when we get it. GBIF requires us to support all of this information. Having this information will enable us to quickly generate lists of data providers and data owners that must be cited in a publication, and will enable us to "link back" from publications to their data providers, to produce a list of publications citing a given source within a given years. Herbaria in particular are particularly keen on getting this information in exchange for use of their data.

Changes to the BIEN 3 core schema will be required to support metadata pertaining to data provenance, ownership, attribution and access conditions.

*Scope of implementation*: Source-specific. These operations must be performed at the loading stage using rules specific to individual data providers. However, to be useful, source-specific methods must be mapped to universal sets of method descriptors.