**BIEN3 Derived Data Products**

Prepared by: Brad Boyle

Date: 6 Mar. 2013

Last revision: 6 Mar. 2013

**Contents**

## Introduction

This document describes derived data products that will be provided as part of the completed BIEN3 database. "Derived data products" are not present in the raw data, but are calculated based on those raw data with input from additional data streams.

## Derived data products

### Taxonomy

##### Taxonomic name resolution

* Name correction using TNRS
* Matched name and accepted name
* Original (verbatim) name also preserved
* Important to store taxonomic source(s), date of access of TNRS, versions of TNRS and TNRS database

### Phylogeny

##### Major higher taxa

* Join taxa in observations to nodes within the NCBI classification
* Prepopulate first-level elements (columns) representing key higher taxa
* Necessary for distinguishing higher taxa such as mosses, lycophytes, ferns, conifers, angiosperms, etc.

##### BIEN Phylogeny

* All taxa in BIEN joined to species-level molecular phylogeny
* Variety of assumptions involved for BIEN species without sequence data
* Linnaean taxonomic categories used to splice in these taxa
* BIEN-iPlant collaboration

### Geography

##### Geographic name resolution

* Discovers and standardizes names of political divisions within 3-level hierarchy (e.g., Country, State, County)
* Standardize and index names in **both** BIEN observations and GADM political division shapefiles needed for geovalidation
* Required first step for geovalidation
* Ultimate goal: "Geographic Name Resolution Service" (GNRS), publicly accessible web service, analogous to TNRS
* Products: IDs and standardized names of three political divisions
* Status:
  + BIEN2 prototype implemented by Brad and John
  + BIEN3 prototype completed by Jim
  + Waiting to be implemented by Aaron
  + Possibly add fuzzy search?

##### Geovalidation

* Verify that declared political divisions fall within known limits of that political division
* Requires GADM shapefiles
* Require geographic name resolution as first step (see above)
* Product: Column isGeovalid (0,1,NULL)
* Status:
  + BIEN2 prototype implemented by Brad and John
  + BIEN3 prototype completed by Jim
  + Waiting to be implemented by Aaron

##### New World localities

* Flag New World vs. Old World localities
* Product: column `isNewWorld` (0.1)
* Status: completed

### Disturbance

##### Cultivated specimens

* Detect and flag specimens representing collection of cultivated (planted) plants
* Inputs:
  + Indications of cultivated status from original data provider (source specific)
  + Key words in locality description (e.g., "garden", "plantation", etc.)
  + Proximity to herbarium (currently use 3 km)

##### Cultivated, planted and otherwise disturbed plots

* Currently applied only to FIA plots
* Any plots detected as representing plantations or logging treatments are excluded prior to loading
* Question: should we be this strict? Shouldn't we merely flag instead of exclude? But, if flag, will need controlled vocabulary for indicating types and degrees of disturbance
* Plots representing secondary/successional vegetation are important for some research questions
* We have several such successional series in SALVIAS, now in BIEN

### Conservation status

##### IUCN Threatened and endangered status

* Status transferred directly from latest IUCN Red List

##### BIEN estimated candidate IUCN status

* Indicates species possibly eligible for IUCN T&E status, based on range size and fragmentation
* More comprehensive estimate of IUCN status for species with abundance data as well?

##### Other sources of T&E status

* Should we consider linking to other T&E lists (country, state, regional, other NGOs)

### Demography

##### Local abundance in plots

* Summarized abundance at plot level for various types of plots
* Knowledge of plot sampling metadata critical:
  + Abundance need to be scaled to standard area
  + Plots using different inclusion/exclusion criteria **cannot** be compared, without additional subsampling.
  + Even subsampling may not be possible in some cases

### Sampling methods

##### Plot size and shape

##### Inclusion & exclusion criteria

##### Subplots

##### Strata

Standard methodologies

* Index and name sets of methods (plot size, strata, inclusions/exclusions) representing "standard" methodologies (e.g., "1 ha/10 cm dbh tree plot", "Gentry plot", "Braun-Blanque Releve")

### Data ownership, attribution and access

##### Data indexers, data providers, data owners

* Must extract complete information

##### Mechanisms for data attribution

* Critical for compliance with GBIF MOU and BIEN mission

##### Data access conditions

* Will require web user database and interface to implement

### Inferred traits

##### Growth form

##### Phenology

##### Abundance

##### Size: Height

##### Size: Stem diameter

##### Geographic distributions

* Range maps
* SDM statistics