**BIEN 3.0 requirements outline**

* + Core DB (by Jan. 2012)
    - Redesign VegB (by mid-Nov)
    - People: BP, ML, NS, BB, SD
    - Walk-through of VegB model (Aaron & ML)
    - Test-loading of samples of specimen and plot data [Brad, Bob, Steve: provide example data sets]
      * Specimens: NY
      * Plots: To be determined (Brad, Bob: today!)
    - Check against DWc to make sure all key elements supported (early Nov.)
    - Identify issues (mid Nov.)
    - Are there elements that should be REPLACED by DWc elements
    - Agree upon revised data model (early Dec.)
    - Generate & test database
    - BIEN-extension DWc (early Dec.)
    - People: BB, BT
    - Revised DWc-BIEN (early Dec.)
    - Feedback to TDWG (MS,BB,NS) – separate effort; should not delay BIEN3.0 development
    - Redesign VegX (end Jan.)
    - People: NS, with BB
    - Preliminary revision based on issues raised at BIEN meeting (end Jan. 2012)
    - Include support for vouchers (biological collections, housed in herbarium) as DWc objects.
    - Evaluate mapping with respect to DWc and VegB
    - Other issues:
    - GUIDs
  + Mappings from schemas to core DB (Jan-Mar)
  + DWc🡪VegB (Aaron, with assistance from ML, BB, others?)
  + VegX🡪VegB (Aaron, with assistance from ML, NS)
  + Get data
  + Bien2 data (end Jan.)
    - Use new raw data where possible
    - Specimens: Where possible get directly from existing DiGIR/TAPIR servers
    - Otherwise use old raw data
    - Acquire all missing metadata:
      * BP: VegBank, CVS
      * Rick: CTFS
      * FIA (BB)
      * Other
  + New data on polygon occurrence (BP)
    - Error metadata required
  + New data (Made all contacts by end Feb., deadline for acquiting data end Apr.)
    - Identify new sources (BB, BE, BM, BT, BP, PJ, others)
    - Search GIVD for additional new word databases & contact owners (BP)
  + Load data
  + Development of plot🡪VegX mapping tool (check with Mark, Jim regarding timing)
    - Possible development of open source by Aaron mapping tool based on existing applications:
      * VegBranch (work with ML)
      * NVS data mapping tool (work with NS)
    - Publication (BB to lead)
  + Specimen data🡪DWc: Aaron work with BB (Mar.)
  + Plot data🡪VegX: Aaron with NS, with other people with knowledge of specific plot data (Apr-June)
  + Testing of loaded core db (June/July)
  + Validation
    - Integration and optimization of existing validation scripts
    - New validations:
      * Georeferencing
      * List of others
    - Where does this happen?
      * Point of data upload?
      * Loading to analytical model?
    - Feedback to data providers
      * Geographic scrubbing results
      * Taxonomic scrubbing results
      * Other validation (data type, domain consistency)
  + Derived data products
    - Range from raw data to highly-derived analytical products such as range maps
    - Why control data products:
      * Versioning
      * Performance
      * Convenience
      * Repeatability
      * Simplify data end-product distribution
      * Create new/more information
    - List of common data products
    - External products?
      * Phylogeny/systematics
      * Traits (currently in BIEN, but could “outsource” to traits group)
      * Climate/geospatial
  + Public access point
    - What do we want to get out (use cases)
    - Versioning
      * Final decision to be made later; development should support all options
      * Regularly scheduled refreshes of EACH end-point data product
      * BP: access to two end products: mirror of core db, changes minute-by-minute; separate analytical products, refreshed less frequently, also versioned
      * Version the end-point, even if database refreshed more frequently
      * Refresh frequency to be decided
    - API
    - User interface
      * Data discovery and download
      * Essential, but details to be decided
      * Via current BIEN website?
      * New website?
      * Not certain, to be decided
      * Data entry/editing/management tool
      * Peter: will generate more interest from wider community
      * – out of scope; but potential to develop collaboratively with outside support from interested stakeholders (e.g., CI-TEAM)
      * User data upload tool – out of scope; possibility to develop as part of preceding data entry tool

**Development procedure**

* Directed by Jim or Mark (to be decided)
* Planning meetings/web conferences
* Minimum once per month

**Database infrastructure requirements**

* Automation:
  + To maximum extent possible
  + A complete data pipeline
  + At all stages: data acquisition, validation, end-point publishing
* Validation:
  + Must faithfully preserve or extend the functionality of current validation scripts
  + Preserve original values for comparion with validation results and aid with improvement of validation
  + Validation results should be captured in logfile and returned to original data provider

**Publications**

* VegX – done!
* Comparison of architectures of BIEN2.0 vs. BIEN3.0 (SD)
* BIEN3.0 White Paper (all of BIEN, lead by BJE)
* VegX mapping tool paper (BB, Aaron, NS, others)

**Data ingest requirements**

* Make sure following are all accommodated:
  + Specimens: all DWc elements, plus BIEN extensions
  + Plots: all essential elements (with reference to VegX and raw data)
  + Species observations (e.g., county-level occurrences)

**Data end-products requirements**

* Plot identifiers (“community”;=plotID, plotCode (if unique))
* Tiered access control
* Provenance (dataset, data owner; for attribution & acknowledgement)
  + Critial end-point requirement
  + With every download, user should receive file containing:
    - Who owns the data
    - Data owner contact information
    - Data use conditions, if any
    - Proximate provider (e.g., “Cyrille Violle”, “Missouri Botanical Garden”, “GBIF”)
    - Primary data provider (if proximate is an aggregator)
* Taxon
  + Species
  + Other taxonomic level: Family, genus, infraspecific taxa
  + Major higher taxa
    - Filtering homonyms
    - Restricting analyses to particular groups (such as flowering plants, conifers, not ferns, etc.)
  + Phylogeny:
    - BIEN species mapped onto TOL
    - Needed for selecting or excluding major clades
* Locality
  + Latitude, longitude
  + Political divisions
  + Detailed locality descriptions
  + Needed for some validations; esp. cultivated/exotic specimens
* Specimen data:
  + Detailed specimen descriptions
    - Needed for some validations: cultivated/exotic specimens
    - Useful to extracting traits such as habit, growth form
  + “Cultivated” flag, metadata on reason for flagging
    - If provided by original source
* Plots
  + Abundance
    - In plots (plot metadata critical for proper use)
  + Size measures from plots (e.g., DBH, height, etc.)
  + Full method metadata
    - Area, layout, unit of observation (species vs. individuals), size and habit cutoffs
* Traits
  + Should include key metadata, including how collected, original citation if provided, data owner
  + Unit, method
* Climate and geospatial data
  + Essential, but out of scope for BIEN core db
  + Critical feeds required for validation and analyses, production of key BIEN data products.
  + For example: DEM needed to populate elevation field where missing
  + Climate layers for range models
* Species diversity
  + Alpha (local) diversity from plots
  + Meso-scale diversity
    - In grid cells via aggregating from plots, specimens
* Species pool selection:
  + Ability to get list of species pool for a particular point or polygon
  + List based on all species with modeled ranges intersecting that point
  + List based on all species with observations intersecting that point
* Species occurrence from plots, specimens, traits, political divisions
* Climate layers
  + Out of scope? Unless use climate data provided with some plots
* Traits (seed mass, sla), aggregate by both genus & species
* Species ranges
  + Georeference point occurrences
  + Species occurences by polygon (political divisions)
  + Modeled ranges (various methods)
  + Expert maps? Where available as feedback from BIEN projects? Links to MOL?
  + Derived calculations:
    - Niche tolerance values (e.g., typical output from Maxent)