Potential validations on BIEN data

Contributions/editing by Susan Wiser, Nick Spencer, Barbara Thiers, needs input/comment from Brad, Mark, Bob

Working group meeting 2012

**Types of Tests to conduct**

1. **BIEN 3.0 Data load pipeline functioning as expected**
   1. **Can BIEN3 construct the view of the data as provided by each Source (or flattened views of such)**. This needs to encompass all fields migrated from Source data into BIEN3. The desired format for checking needs to be specified by the Source data provider. Only a small number of records need to be checked (say 10). This was done at meeting for NYBG and MO data; is in progress to be done for SALVIAS and VEGBANK.
   2. **Consistency in distribution of data values in source versus BIEN3**, i.e. for categorical fields (e.g. soil drainage class, bedrock type), all distinct categories that were present the source need to be present in BIEN3 and the frequency distribution of observations in these categories also needs to be the same. For quantitative fields (e.g. altitude, slope), statistical summaries (e.g. mean, range) of the source and BIEN3 should produce identical values. Priority fields/elements are those from VegCore that are being passed through to analytical database. Ultimately all appropriate fields/elements in VegCore need to be validated in this way.
      1. ***Specimens & Plots***
      2. ***Specimens only***
      3. ***Plots only***
   3. **Migration audit**: are the numbers of plots, species, trees etc. the same in source versus BIEN3?
      1. ***Specimens & Plots***
         1. Number of species in source= number of species in BIEN3
            1. Need to account for reduction in species number due to TNRS results. Changes made as a results of TNRS should be available as well – Uploader should receive a report of names changes resulting from TNRS [see section below]
      2. ***Specimens only***
         1. Number of specimens in source= number of specimens in BIEN3 same comment as above [see section below]
      3. ***Plots only***
         1. Number of plots in source= number of plots in BIEN3
         2. Number of species by plot consistent [how to account for reduction in species number due to TNRS results]
         3. Number of trees [individuals] by plot consistent
         4. Impacts of application of TNRS: Duplicate rows for a aggregated species importance value in a plot could be created by spelling corrections and/or synonomy corrections. Calculations of aggregate values e.g. stem counts, basal area, % cover need to take this into account]
         5. Calculation results in the analytical database can be replicated by using the source data directly (e.g. basal area, stem density, others??)
   4. **Preservation of data integrity**: is referential integrity retained, do all child attributes have a parent and is it the correct parent?
      1. ***Specimens & Plots***
         1. Are tables/elements that were unique in source dataset still able to be recognised as unique in BIEN3 (i.e. unique values for plotids, reflecting uniqueness at both plot-scale and source dataset scale, tree tags, unique ids for specimens by institution)
         2. Ensure that the uploader /data provider can view the specimen records in DarwinCore format (Taxonomic name, location, collector/collection number, unique identifier, source of data) and the plot records in their original formats. Data provider should verify that the top and bottom xx rows migrated correctly.
      2. ***Specimens only***
      3. ***Plots only***
         1. repeat measures of plots
         2. repeat measures of tagged trees
         3. subplots nested within plots [CVS provides most challenging example]
         4. tiers/strata nested within plots
         5. multiple attributes remain associated with original object (e.g. measured of dbh and height on the same tree retained)
2. **Data errors [source: feedback to data providers, error limits to qualify for upload into BIEN]:** 
   1. **Values out of possible range**, e.g. aspect>360, all categories match authoritative list
      1. ***Specimens & Plots***
         1. Aspect < 361 [where numeric]
         2. Slope < 91 [where degrees], Slope < xx [where %]
         3. Elevation > xxx and < xxxx [where m]
      2. ***Specimens only***
         1. There are data sources that could be marshalled to create a test on collector name and collection date (that is, was the collection made within the life span of the collector?) but this tool does not currently exist.
      3. ***Plots only***
         1. Tree attributes (e.g. dbh, height) out of range. Flag all values beyond 99th percentile.
         2. Basal area as above
         3. Stem density as above
         4. Might want to examine above as ratios i.e. dbh:height; stem density:basal area
   2. **Uniqueness errors [source]:** 
      1. Check for duplicates of uniqueentities, e.g., a stem tag is not duplicated within a [plot|subplot|dataset]**.**
      2. For specimens, there may be duplicate specimens from various data sources. These may have different names because of different determinations. These records should be flagged when uploaded into the analytical database to ensure they are not having disproportionate influence on range determinations. Could be especially critical for species where data are sparse or where the record is on the edge of the known range.
   3. **Location errors [source]:** 
      1. Geoscrubbing where data source only provided lat/long (no location text) country is derived. Country should be cross tabulated by data source
      2. Map of all occurrence data; map of occurrence data by source