Potential validations on BIEN data

**Develop a series of tests**

**Types**

1. **BIEN3.0 Data load pipeline functioning as expected**
   1. **Consistency in distribution of data values in source versus BIEN3**, i.e. all categories present in categorical values in the same proportions, statistical summaries of quantitative values identical. Do a pass through all appropriate elements in VegCore with priority to elements being passed through to analytical database
      1. ***Specimens & Plots***
      2. ***Specimens only***
      3. ***Plots only***
   2. **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.
      2. ***Specimens only***
         1. Number of specimens in source= number of specimens in BIEN3
      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??)
   3. **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. ***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***
      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]:** Check for duplicates of uniqueentities, e.g., a stem tag is not duplicated within a [plot|subplot|dataset]**.**
   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