**Biodiversity Data Cleaning: Ashwin Agrawal-GSoC**

**Date- 30-5-17**

**Fields for DwC summary table:**

Below are the fields which I will include while building the summary table: (**GBIF data**- these are same as discussed in the meeting)

**Taxonomic Fields:**

1. **Name or scientific name:** In the scientific name and name field we can get the frequency of each data entry and therefore get the most abundant species and the least.
2. **Basis of Record:**
3. **Kingdom**
4. **Phylum**
5. **Order**
6. **Family**
7. **Genus**
8. **Class**

For all the above six taxonomic fields namely kingdom, phylum, order, family, genus and class- based on different species and countries the frequency can be calculated. When calculating the frequency I will change one parameter at one time, like if I vary country I will keep species constant and vice –versa.

1. **specificEpithet**

**Spatial Fields:**

1. **decimalLatitude**
2. **decimalLongitude**

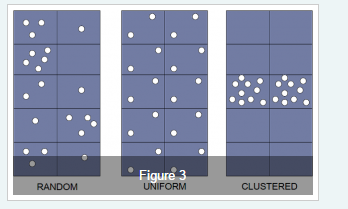
For the coordinates basic summarizing can be done by using measures of central tendency like minimum, maximum, range, mean, standard deviation etc.

Frequency and density of points can be calculated. Spatial descriptive statistics like mean and median centre can be evaluated. These will help in identifying the randomness of the points on map.

Another important parameter for summarizing is arrangement of points in space. Whether the points are random, uniform, clustered or dispersed. This can easily be checked by Complete Spatial Randomness test or simple Poisson distribution of points.

Average nearest neighbour distance can be calculated which will give an estimate of arrangement of points wrt each other.

The area of the polygon formed by the coordinates can be calculated and hence the density of the points can be known. A bounding box around the coordinates can be plotted to know the extent of the area covered by points.



1. **coodinatePrecision**

For coordinate precision a certain cut off limit can be set (up to certain decimal places) for filtering out the coordinates or marking the data quality. Users can vary this limit according to their need.

1. **Coordinateuncertainityinmeters**

Similar approach like that of coordinatePresision can be followed for coordinateuncertainityinmeters

**Geographical Fields:**

1. **EventDate**

The format of the date can be checked and given to user pre-hand, a feature to convert the date format can extended if necessary. Yearly of monthly frequency of species can be summarized.

For the date, year and other temporal fields we can use measures of central tendency where ever applicable and also, we can get density of species with time i.e. between which time period most species were spotted.

1. **Country or country code:**

For country and country code frequency of specific species in specific country can be computed. Similar frequency tables based on different country and species combination can be tabulated. Most common and least common species of different countries can be found.