Linking specimen with material citation and vice versa (Topic 8)

Aim/problem/goal

Material citation cites in scholarly publications specimens or groups of specimens as bases for the research results presented in the article. The most common use of material citations is in taxonomic literature, either as part of a taxonomic treatment or in form of tables, often also including additional links such as genes via accession numbers. Traditionally these citations cite specimens in natural history institutions. However, very few of the natural history collections have an IT infrastructure in place that allows discovering and citing the respective specimens and even more bi-directional linking. At the same time, an increasing number of natural history institutions export on a regular basis their occurrences to GBIF. Interacting with GBIF is an option that circumvents custom solutions for each institution, and at the same time allows the institutions to retrieve the links to the material citations via searching their uploaded occurrences and related, clustered occurrences - that is material citation uploaded by TreatmentBank (TB). Making use of the TB-GBIF interactions allows making use of the linking mechanism by TB which will add the GBIF occurrence ID to the respective TB record, and once concluded re-upload the respective data set including the attributed material citation to GBIF.

Method

We developed an algorithm aiming to link the material citations in the GBIF database and the specimens in the Natural History Museum of Bern (NMBE) collections. The algorithm hinges on calculating similarities between the instances in both sides of linking. It compares each material citation and specimen based on ex-ante selected attributes and calculates pairwise similarities accordingly. The attributes could be in the string type such as genus or family information, as well as numeric type such as latitude or longitude of the discovery place. The algorithm calculates the similarities for each data type separately and merges and normalises them at the end to find a final pairwise similarity in the interval of [0-1] between a material citation and a specimen. It sorts material citations for each specimen according to the similarity score. Finally, it assigns the most similar material citation's "material citation ID" to the corresponding specimen in the NMBE collection. Overall, the algorithm finds the most similar material citation for each specimen; thus, it bridges two datasets.

Results

We developed an algorithm for matching "material citations id" in GBIF to the NMBE specimen.

Conclusion

The use of GBIF as a surrogate of institutional databases has several advantages. First, only one bi-directional linking algorithm and interface has to be developed. The GBIF occurrences are continually, automatically updated, whenever an attribute has been added on the TreatmentBank side. The institutions are at ease when they want to update their records. The development of the clustering algorithm (see also Topic 3) will facilitate linking specimens and material citations from a particular institution in a first step - the searching over a billion occurrences will be very time consuming and not doable for linking large numbers of material citations. This is even more complex by the nature of material citations that can represent the entire specimen record to only parts, often in slightly different formats.