Annex 4 - Biodiversity Enhanced Location Services (BELS)

In order to make the georeferences produced during the course of the project easily and openly available, we developed a gazetteer of all distinct combinations of terms of the Darwin Core Location class coming from data aggregated in GBIF and iDigBio using snapshot of both sources from January 2021. All Locations georeferenced during the CESP project are in the process of being added to the gazetteer as well. This gazetteer forms an integral part of the Biodiversity Enhanced Location Services (BELS) discussed in the community forum "Imagining a global gazetteer of georeferences" (https://www.idigbio.org/content/darwin-core-hour-2-bbqs-imagining-global-gazetteer-georeferences).

The gazetteer contains preprocessed determinations of simplified representations of all Locations in strings for matching one Location with another. The simplifications are to increase the capacity to match two Locations that are the same place even if there are variations in how they were captured textually. Examples of simplifications include a) the normalization of distinct Unicode characters that have the same meaning to a single common value, b) removal of punctuation that does not affect numbers, c) setting all matching strings to lowercase, and d) the translation of accented characters to ASCII equivalents.

In addition, the gazetteer contains preprocessed determinations of the best existing spatial representation, if any, for every matching string. The majority of matching strings have one or more spatial representations that include at least coordinates. The best spatial representation among those existing for a given matching string are determined by their compliance with georeferencing best practices (Chapman & Wieczorek 2020, Zermoglio et al. 2020) as well as by their ability to represent the whole set of possible spatial representations for that string. The technical details for this and all other aspects of the gazetteer construction are the subject of a manuscript in preparation.

To satisfy the goals of the project, we built a set of scripts, an API, and a skeletal user interface (code available at https://github.com/VertNet/bels) that allow users to submit one or more Darwin Core Locations via a CSV file for checking against the BELS gazetteer. Via the scripts, the API, or the web interface, input Locations are processed to produce matching strings in exactly the same way they were made within the gazetteer. These matching strings can then be used to select the best existing georeference from among the Darwin Core Locations that are the same place within the gazetteer. Using the web interface (temporarily at https://localityservice.uc.r.appspot.com/ but eventually destiny for georeferencing.org), one can upload a file with Locations (and any additional data fields) and provide an email address. When the file has been uploaded and processed, the results, which include all of the input fields plus fields for the best georeference available, are saved in a CSV file and the link to the file is sent to the email provided by the requester.

The code for the service is written in Python, while the infrastructure for the service and gazetteer rely on a suite of Google products including App Engine (for the web application), PubSub (for the georeferencing jobs spawned by the web app), Cloud Function (to process the jobs), Cloud Storage (for the uploaded and result files), and the free version of the third party SendGrid libraries (to send the email announcing the finished job and location of the results file). The costs for the development of the gazetteer, API, and web interface were in-kind contributions of the VertNet team to the project. The costs for services (data storage and processing) are supported by the Field Museum of Natural History. As mentioned above, this service is also part of a broader project, recently funded by a US National Science Foundation grant (DBI- 2027241 "Collaborative Research:

CIBR: Leaping the Specimen Digitization Gap: Connecting Novel Tools, Machine Learning and Public Participation to Label Digitization Efforts", affectionately known as ("DigiLeap"), with a start date of January 2021. DigiLeap aims to integrate Citizen Science through Notes from Nature with specimen records and tier images stored in Simbiota and georeferenced through GEOLocate, which in turn will call on BELS to improve georeferencing capabilities. Sustainability for more development and improvement is therefore granted at least for the next four years. It is our hope that the BELS can be integrated in the future with the GBIF infrastructure.