Annex 1 - Georeferencing priorities within the region

Prioritization of location data for the workshops

It was decided that the source of localities and georeferences be the localities already available through the GBIF network. The rationale behind this was that those are localities already published, and that therefore a georeference could directly have an effect on enhancing data quality of records that are already available (as opposed to taking localities from other sources, whose publication is uncertain).

All localities from Argentina and Chile were extracted from GBIF occurrence snapshots (2019-04-15 and 2020-04-08, respectively) and loaded into BigQuery. A location table was built including all distinct combinations of fields in the Darwin Core Location class, where countryCode=AR or CL (for Argentinian or Chilean locations, respectively). Each locality was then assigned a UUID as its localityID, and these were also added to the corresponding Occurrence records, so that georeferences of the distinct localities could be reassociated with the occurrence records. (If needed, these intermediate working tables can be provided, please contact John Wieczorek).

In order to prioritize the localities, a first criterion was that the occurrences associated with the localities be from Argentinian or Chilean data providers. This criterion was adopted so that the repatriation of the georeferenced data could be more easily achieved. In order to filter by provider, a secondary table was built with distinct combinations of institutionCode -collectionCode associated with the occurrences that had countryCode=AR or CL.This list was manually processed to determine which institutions belonged to the target countries.

Data from some providers were purposely left out given that:

a) it is known that they publish records from other local institutions that also publish,
b) difficulty was identified in repatriating enhanced data further along the line, e.g., data coming from eBird, and
c) it was impossible to easily determine the origin of the data from institution and collection codes (e.g., two letter codes).

Aside from eBird data, the rest did not represent a large proportion of localities nor of occurrence records associated. With the information about providers, an extract was obtained of the localities from Argentina or Chile associated with occurrence records published by institutions in those countries.

The next prioritization step was based on the number of occurrence records associated with each locality, ordering from most to least represented locations. For **Argentina**, the first 5550 most represented locations

were distributed among the students for georeferencing (about 150 locations each). This number of localities greatly exceeds the expected number of localities to be georeferenced during this project. The number was chosen contemplating that not all students would finish their task and that there may be localities that cannot be georeferenced due to lack of sufficient data or to internal inconsistencies. The distribution of localities among students was not random, it was prioritized according to the regional provenance of each student and their knowledge of each region. Participants were asked to tell which provinces they felt more comfortable working with, and hence selections by dwc:stateProvince were made as a first pass to assign the localities. For those that had the field empty or not standardized, a pseudo-random assignment was performed. Prioritization for **Chile** localities followed the same criteria expressed above, and 110 locations were designated to each participant.

Prioritization for future projects

Based on the methods applied above for prioritizing the localities to georeference during the workshops, we could identify the gaps in the geographic data, and their quality, for both countries.

First, we observed that, although relatively few locations (i.e., a few thousands) accommodate most of the occurrence records, the data shows long tails, with many distinct localities applicable to single occurrence records. Considering that all Darwin Core Location class terms are included to determine if a locality is different from another one, slightly different locality strings that may refer to the same actual place are considered as distinct localities. For example, a location where dwc:locality="12.6km S Antofagasta" is different from one where dwc:locality="12.6km S from Antofagasta", even though a human interpretation could match the two. This conservative approach ensures that no assumptions are made regarding different strings. However, we recognize that a standardization step prior to the processing and georeferencing may render the workflow more effective. Standardization may include format normalization (e.g., punctuation, capital letters, etc.) but also cleaning steps to homogenize versions of a given place name (e.g., abbreviations, full names, etc.). This may be a difficult task to perform in some cases, and would greatly benefit from local knowledge.

Second, we observed that many locations have only coordinates -and no locality textual description. Georeferencing localities of the type "only coordinates" is the simplest case (Zermoglio et al. 2020), and would also be the one that could be automated more easily. We are currently in the process of writing a script that can interact with the Georeferencing Calculator (Wieczorek & Wieczorek 2019) to get those georeferences without human intervention.

Third, and similarly to the previous case, many locations are only described to administrative levels ("higher geography" levels, e.g., only to province, or county). These cases could also be subject to

automatic georeferencing by assigning coordinates and associated uncertainties of the given administrative divisions. The BELS project, which will serve as a continuation of this project (see Annex 4) contemplates this automation.

With the above, if standardization and automatic georeferencing was achieved before hand, more records could be completed and the effort dedicated to the manual georeferencing process could be instead allocated to disambiguate difficult localities and/or to check for final data quality. Therefore, both processes (preprocessing standardization and automation) should be considered as a priority for future georeferencing projects.

The workshops held during this project provided an additional opportunity for the project partners to assess the priorities of the local communities. Among these, we identified the following areas of interest for future georeferencing projects:

- ***Marine locations***. Both our countries have an extensive coastline, and many records from marine locations. Some localities from these records consist of only coordinates, with very little other location information. As mentioned above, these could be relatively easy to georeference if an automation process was in place. Other marine localities consist of textual descriptions, usually referencing coastal entities. This kind of localities present a much bigger challenge and georeferencing them often requires detailed interpretations. Furthermore, disambiguation for these types of strings may demand working on an occurrence record by record fashion, where other types of data are considered (e.g., taxonomic information). In the future, we believe it would be advantageous to coordinate projects that exclusively deal with marine locations, and that attempt to align efforts with other initiatives and organizations (e.g., OBIS).

- ***Protected areas.*** Data from protected areas, both historical records and data from more recent and current monitoring programmes, are of particular interest to assess the effectiveness of conservation efforts. Locality descriptions from protected areas often include pieces of information about specific places within those areas, with designations that are internal to the corresponding administration agencies. Teaming up with such agencies in future projects would improve interpretation of these localities and at the same time serve as opportunities to train personnel that are constantly recording location information and feeding it into the GBIF network.

- ***Historical records***. Historical records are intrinsically challenging but they constitute a fundamental resource to set the baseline for understanding biodiversity trends. Location descriptions in this kind of records seldom have coordinates associated, and often contain old denominations for administrative divisions and places, no longer in use and difficult to find. Furthermore, in the past there was probably much less awareness of best practices for capturing

location information and the importance of associating it to individual records. For instance, great sources of location information are the collectors' field notes, but more often than not these are not directly linked to specimens labels or ledger entries. To add value to historical records, and particularly regarding the quality of georeferencing data, projects in the future may focus on integrating other types of information, for instance cross-linking the location information available on a record-by-record basis with collectors' itineraries and field notes descriptions. These kinds of projects may benefit from using techniques that expedite digitization of these resources using (semi)automated approaches, such as imaging and use of optical character recognition (OCR) algorithms.

- ***Observation records***. Currently, observation data constitutes over 85% of all records shared through the GBIF network. For our region, although with some variability, observation records are also predominant. The bulk of these records are generated through citizen science projects (e.g., eBird and iNaturalist) and contain coordinates, but little other geographic information. Some of the platforms through which these data are generated do include other data, but these are not necessarily shared together with the observations. While these records are relatively easy to georeference based on coordinates only, repatriation of those georeferences is challenging, and would require a concerted effort including those citizen science project managers. Such collaborations would have the potential to improve a very large amount of records, and could be approached from national, regional or global perspectives.