BiNHum Harvesting and Indexing Tool

Documentation

SHA algorithm: should a harvested data be processed

In order to increase the performance for processing the previously indexed dataset, a new algorithm has been developed, based on a checksum comparison of the downloaded files. Once the records have been harvested and saved locally in zipped files, B-HIT calculates a basic checksum of the downloaded files, using the SHA algorithm[1].

If it is the first time the dataset is harvested, then each file will have to be processed and the newly calculated sum will be saved in the database along the filename, the dataset name, and the type of harvesting (default/associated/sibling/extra, see details below). If the dataset has already been harvested in the past, 4 distinct use cases can be distinguished:

1. The new checksum does not differ from the old one saved in the database: there is a high probability that the file content did not change since the last time: it can be assumed that there is nothing to be updated in the current database.
2. The new checksum differs from the old one saved in the database: the old records have to be deleted from the database and the newly harvested file has to be processed. The checksum in the database will be updated with the new sum.
3. The new downloaded file did not exist previously: it means that the dataset contains new records or that the units were dispatched in another order because of some adjustments (i.e. deletion) on provider side. The records will be considered as new and will be harvested and added to the database.
4. Some files listed in the database were not returned by the dataset: some units have been deleted on provider side or may have been returned within another bundle of units (see point 3). In that case, the corresponding “old” records will be removed from the database.

Quality tests

The first test consists in translating the country name in English. The different accepted input values are extracted from the multiple language files and are completed based on the most common errors and typos met (i.e. “Iatlien” instead of “Italien”, missing empty spaces etc.). States and regions are also commonly mapped as countries by the providers: their affiliation has also been added to the known inputs. The result of this translation process is saved and can take four values:

- OK (it could be translated easily or it was already in English),

- CORRECTED (it could be translated but the original value was not a standard translation),
- FAILED (no translation could be made)
- N/A (empty input, the tests could not be run).

If the original country does no longer exists or if it borders do not fit any actual country, the corrected value will be set as an “Unknown or unspecified continent” (continent being replaced by Europe, Eurasia, Asia, North and Central America, South America, Oceania or Antarctica). “Unknown or unspecified country” will be inserted in the database for the empty values or for the characterisable values.

The second test consists in trying to extract the country based on the locality and the gathering areas.

The third test consists in standardising the ISO code in its 2-letter standard. The original value can be correct or non-available, can be replaced (ISO no more existing, i.e. SU) or corrected (3-letter code to 2-letter code).

The forth test consists in comparing the validated ISO and the validated country name. During this comparison, if no coordinates are available, a missing value will be inferred from the available country or from the available ISO (i.e. ISO-code=ZZ (unknown) and country name=Germany, coordinates empty-> ISO-code inferred as “DE”). Inferring a value will lead to a warning information in the database. The comparison can be successful (returns OK), can fail (ISO and country name do not fit together), can return N/A if one value was incomplete or unknown.

The fifth test will check the coordinates validity – the text values are parsed into decimal values, their ranges are checked (i.e. is the latitude between -90 and +90, resp. -180 and + 180 for the longitude).

The sixth test will compare the cleaned country data and the coordinates, using different methods and services. If the system previously detected an inconsistency between the ISO-code and the country name, it will try to improve the data based on the coordinates. For example, the units ISO-code is set as “DE”, whereas the country is set to “France”, and the coordinates are “47.632082, 7.560282”. The 3rd test “failed” for this entry. The current test reveals that this coordinates are supposed to be in France. As “France” belonged to the original values, the system will infer that the country name was correct and the ISO-code was erroneous. Thus, it will replace “DE” by “FR” in the second database table. If the coordinates lead to a third value (ie. Germany, France, and Switzerland), the quality job will save that specific information but won't alter the values in the database, as there is no way to figure out which one is the correct one. If the ISO-code or the country name was missing and the coordinates are consistent with the existing country indication, the missing field will be inferred.

If the original values of the coordinates do not fit the country fields, the quality process will check the opposites values of the coordinates (+latitude +longitude; +latitude -longitude; -latitude +longitude; -latitude -longitude), add a leading number (many latitude were truncated and missed the first digit), and exchange the latitude with the longitude.

Tests on coordinates reuse existing code from Gisgraphy, KDTree and Geonames.

Eventually, the gathering dates are checked and converted into the YYY-MM-DD format, and the gathering year is extracted.