Harmonizing place names

Harmonizing place names

Last data file 2025-01-03, last updated processing 2025-01-16.

Catalogue data entry practices typically aim for precision. This means that publisher names and locations are entered as originally printed, where substantial variation in presentation can be present. In a long-spanning dataset this can include historical variants of placenames (e.g. Dorpat, Tarbatu, Derpt, Jurjev, Tartu), grammar (Tartus, Tartun, In Tartu), localized versions of placenames (Tartto, Tērbata,), historical spelling variants (Tarto-linan, Tartolinan, Iourieff, Jujew). The entries often have additional specifications too, including other variants (Tartu – Dorpat, Youriev (Tartu)), wider geographical areas (Tartu (Tartumaa)), markers for other places of publication (etc, jne), publisher name placeholders (Tartu:[s.n]) and even several placenames sometimes (Tartu [i.e. Torino], Tallinn [p.o. Tartu]). These variants present a challenge for data analysis, as the data field needs to be harmonized in aggregating all books that were, for example, published in Tartu.

Data harmonization is a common task for adapting catalogue data for computational analysis, where a general framework suitable for any dataset is difficult to offer. The historical variants and spelling variants cannot effectively be derived from the placenames, with even predictable variants relying on language-specific patterns. Additionally, the extent of variation within a particular dataset can be difficult to predict and will inevitably rely on custom effort in each case. There are some general patterns: for example, sometimes the placenames include “Printed in”, or “etc”, sometimes they rely on local grammar, books printed in another language are likely to include a localized version of the placename, however even their adaptation is language and case dependent.

To harmonize the placenames, we built a general workflow that relies on external geographical databases that contain the relevant placenames as well as their historical variants. To do this, we first removed the grammatical markers from the placenames where possible. In Estonian case, the frequent forms here were placenames ending with an s, which were extracted from the set and manually annotated for nominative variants of the names. We then used several geocoding services to link the placenames to geographical coordinates: ArcGIS, Geonames and Google. These geocoding services rely on both historical placenames and fuzzy matching to find the coordinates and build on different databases for the matches. Additionally, we adapted a Geonames dump to find coordinates for the names with a preference for locations in or near Estonia and locations with larger populations. If all the coordinates found were within 20km from each other, we considered it a correct match and relied on it. If the placename was matched with coordinates more distant from each other, indicating that the databases contained different priority entries with that name, we manually resolved the location. Based on these comparisons we iteratively constructed a list of rules and exceptions that we then applied in a rerun at the beginning of the workflow. As an additional control, we were able to rely on country-codes given in the MARC data format. This country-code list in MARC format has gradually been developed since its inception in 1967, with the last update in 2003. As a result, they don’t fully match modern country borders, however, they can be well used to compare with the general area in question. These country-codes were given by the cataloguer demonstrating also some variation in itself. For placenames with conflicting coordinates we checked whether any of the coordinates were within the country marked in the code and preferred these coordinates if this was the case. Some books had several publication locations represented with a semicolon: in this case all placenames were processed separately.

In all, there were 5254 place name variants in the dataset. We harmonized 4545 of these variants to match 2471 harmonized placenames, while applying 405 rules and exceptions based on regular expressions and variant matches to clean the dataset. After applying these methods, there were a total of 4250 unique placenames in the dataset. 4300 of them could be geocoded with ArcGIS, 3883 of them with Google, 901 of them with Geonames, and 2294 in working with a local copy of Geonames with historical placenames. 4467 had all the acquired coordinates within approximately 20 km from within each other and were considered non-conflicting matches. 787 placenames showed bigger differences. In this case, if any of the placenames were in or near Estonia, they were preferred, providing a solution for 407 places with no conflicts on coordinates in Estonia. Of these, 592 conflicting placenames were then resolved manually, in preferring the geographic match that was in or nearby the area marked in the MARC data format. 174 placenames could not be reasonably resolved to particular coordinates with this method. Through the process 70,980 books were given a new name that was either cleaned up or grouped together with other name variants.

As a result, 4533 unique placenames in the dataset have been given coordinates, providing a coordinate for 99% of the books (n = 312,002), 174 placenames there have not been resolved, and 1220 books in the dataset do not include information on the place of publication. Harmonizing place names enabled more accurate geocoding results for non-standard spellings (coordinates were added or corrected for 298 place names, or 6% of the set). The harmonization and linking of the places of publication is likely to contain some errors, but a manual verification showed the results to be reasonably accurate (96% of 200 randomly selected placenames showed accurate coordinates).