Automated Geo-resolution of Place Names in Historical Serial Sources

Erik Radisch[1]

Abstract: This paper is a presentation of the historic place name locator[2], an algorithm, which provides a solution for an automated geo-resolution of place names in historical serial sources. It proposes an approach, which takes historical boundaries into account and can handle variations in writing. This approach can greatly contribute not only to save the content of a source in a database but access its historical meaning.

Keywords: Historical Place Names; Geo-resolution; GIS; Metadata Enrichment

1 Introduction

One of the biggest challenges in building and maintaining semantic databases for historical topics is not only to store the content of the sources, but also to make their meaning accessible. This problem is also present in other semantic databases, yet it takes on a larger dimension with historical topics. Different spellings, the use of terms that are outdated and the loss of „domain knowledge“(place names that are forgotten) make the accessibility of meaning much more difficult. This paper presents an approach which dwell on how the access to an important information carrier in historical serial sources – place names or to be more precise the geo-resolution of them – can be automated.

Automated geo-resolution of historic place names had been already addressed in several approaches. Yet, those approaches were either highly specialized for very particular problems, which cannot be generalized like for example the solution of Schürer et al [SPS15]. Their solution for automated geo-resolution is highly convincing, yet was particularly coded for their very specific three-level-place-name source (parish, county and country). Or they sacrificed potential domain knowledge for the generalization of the algorithm like in the case of the Edinburgh Geoparser. This geoparser gives only rudimentary possibilities to include a historic context of place names in form of a bounding box (a bounding box of the German Empire would include hole Bohemia and a large part of Poland as well).

[1] Saxon Academy of Sciences and Humanities in Leipzig, Germany radisch@saw-leipzig.de
[2] https://github.com/erikradisch/historic-place-name-locator

This paper presents an algorithm which tries to overcome these limitations. The historic place name locator[3] is an algorithm, which enables automated geo-resolution of large numbers of place names. This has been achieved by three important main features of the program: A collated gazetteer, a custom search algorithm and the inclusion of the historical context.

2 Gazetteers

There is no doubt, that the choice of the gazetteer can fundamentally influence the search result. Search results can always be only as good as the chosen gazetteers was. There are some global gazetteers such as Geonames, which reach an impressive coverage, yet they do have often gaps concerning deserted towns or historic names. For this, specialized gazetteers like the historic gazetteer could over better coverage, yet such gazetteers do not even closely reach the coverage of global ones. Thus if one wants to reach high generalizability, it is inevitable to combine several different gazetteers. The historic place name locator combines several different place name gazetteers. The most important are: Geonames, the historic Gazetteer (GOV), Wikidata and Open Street Map (OSM)[4].

3 Search Algorithm

Historic place names do often appear in sources with spelling variations. Those outdated name variations do have a very little chance to appear in current gazetteers. Thus it is very important to implement a search algorithm, which can also deal with spelling variations. In the historic place name locator, a complex search routine consisting of three different similarity search algorithms is implemented. Two of them are the approximate string matching algorithms Damerau-Levenshtein [Da64] and Jaro-Winkler-Distances [Wi90]. Third one is a phonetic algorithm. Here, the user can choose between the Cologne Phonetics [Po69] and Double Metaphone [Ph00]. Those three different algorithms enable the historic place name locator to even out spelling differences between historic sources and current gazetteers.

[3] The algorithm and a manual how to use it can be found under the following url: `https://github.com/erikradisch/historic-place-name-locator`

[4] So far: Wikidata: `https://www.wikidata.org/` [retrieved 07.01.2019], the Historic Gazetteer: `www.gov.genealogy.net` [retrieved 01.06.2019], Geonames: `https://www.geonames.org/` [retrieved 20.03.2019] and Osmnames, a gazetteer based on Open Street Map: `https://osmnames.org/` [retrieved 01.10.2019]. The combination of those different databases to one common is documented here: `https://github.com/erikradisch/historic-place-name-locator/tree/master/make-place-name-db`

4 Historical Context

The last feature of the historic place name locator is its ability to include historic boundaries within the search. Search within its historical context as already performed by the two examples of previous geo-resolution algorithms. Yet, Schürers algorithms depends heavily from a corresponding gazetteer, with this additional metadata. The overwhelming majority of historic projects do not have this advantage. The Edinburgh Geoparser on the other hand does provide the possibility to focus the search on a special bounding box, yet those are very inaccurate. The historic place name locator solves this problem by including shape files of historic place names within the constructed gazetteer. There is a constantly growing number of professional shape files of historic boundaries, which are available under open access. Some examples might be a map of all regions of Europe around 1900 (Mosaic[5]), the borders of the states of the German Empire (Mosaic, Harvard Geospatial Library[6]), the Empire and Kingdom of Austria-Hungary (Mosaic) and the Russian Tsar's Empire (Ristat)[7].

The user only needs to connect the historic place names to the corresponding region in the shape file by providing a second column with the naming of the regions from the shape file (needless to say, that this step gets unfortunately labor intensive, if a lot of places have different historic contexts). The algorithm than favors results from this region. If the algorithm did not find a place in the historic region, it is possible to expand the search area step by step. For example, if a place might not be found in a historic region, for example Hessen-Nassau, a user can then let the algorithm search only in the German Empire and only in a third step in the whole world.

Including a historical context can help to boost the accuracy of georeferencing historic place names greatly as it helps to exclude possible hits which are more unlikely due to their location. As historic boundaries are very often highly complex an automated search algorithm has a real advantage here, as it is often hard to say for humans, where exactly a historic region ended and another one began.

[5] `https://ehps-net.eu/databases/mosaic-project` [retrieved 10.11.2019]
[6] `http://hgl.harvard.edu:8080/opengeoportal/` [retrieved 20.03.2019]
[7] `https://ristat.org/` [retrieved 10.05.2019]

5 Evaluation mode

The program has also an implemented evaluation mode, which enables the user to compare the matches to a gold standard. The algorithm produces HTML files of differing results on which a map is seen with the historic border (if given), the gold standard (green) and the result of the algorithm (red) as can be seen in figure 1.

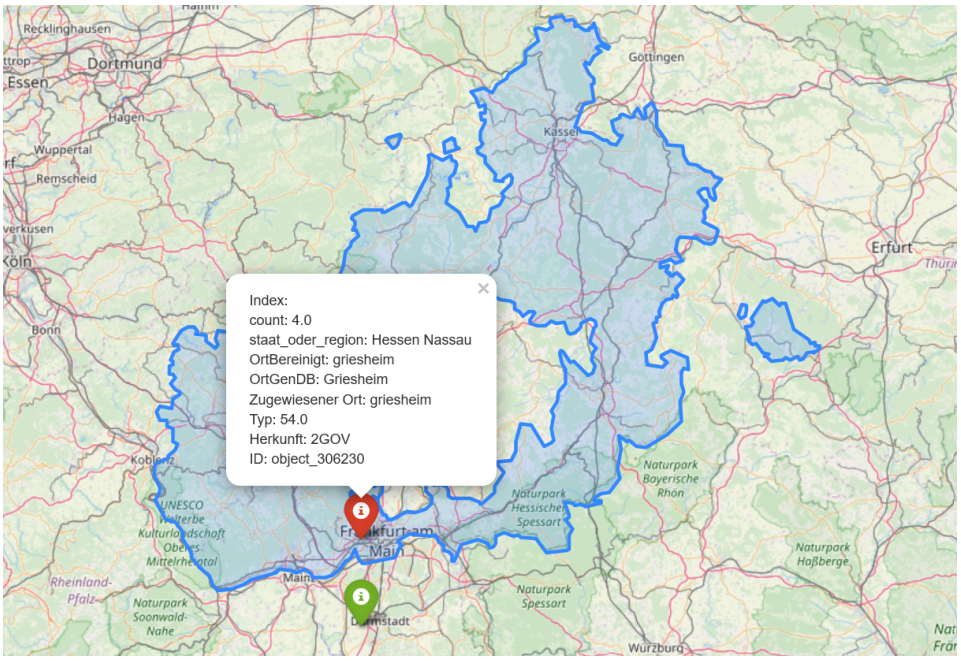

Fig. 1: A sample of the output of differing results. Note that the algorithm might have been right in this case as the hit is directly within the borders of Hessen-Nassau while the gold standard is only close by. (Source of the geospatial data: Germany Provincial Boundaries, 1871, German Historical GIS, online linkage: `http://hgl.harvard.edu:8080/HGL/jsp/HGL.jsp?action=VColl&VCollName=GHGIS1914PROVINCES`; Open Street Map, online linkage: `https://www.openstreetmap.org`)

6 Conclusion

The historic place name locator still demands a considerable amount of preprocessing. A cleaning of the place names might still be necessary. Also the historic context of the place names has to be assigned to a polygon in a shape file. Nevertheless, the historic place name locator offers a generalizable solution for the geo-resolution of place names in serial sources, which also considers the exact historical context. By including the historical context of the place names, the algorithm achieves considerable accuracy. Several tests on a gold standard of around 500 human located place names of several different sources produced

F-Scores close to 0.9. However, the good results should not distract from the fact that some problems remained despite the high allocation rate. The step-by-step search for example can also produce problematic hits. A case of this error susceptibility can be provided by the following example: A list of place names is given, which do have a historical context of several different regions of the German Empire. To keep the error rate low, it might be a good idea to include a second step search within the borders of the German Empire, if a place was not found within the borders of the assigned historical context. If a search algorithm looks first in a part of the German Empire, for example East Prussia, and does not find an exact match, it looks in a next step within the borders of the German Empire and might find a match in Breisgau which is close to France. A place with the exact name, very close to East Prussia yet the Russian Empire could have been excluded from the search. A human might consider the place close to Prussia much more plausible than the place close to France. The expansion of the search area in circles or bounding boxes might be more precise. The implementation of such an option is planed but not yet realized.

References

[Da64] Damerau, F.: A technique for computerdetection and correction of spelling errors. Communications of the ACM, 3(7):659–664, 1964.

[Ph00] Phillips, L.: The Double Metaphone Search Algorithm. In: Dr Dobb's, June 1, 2000 `https://www.drdobbs.com/the-double-metaphone-search-algorithm/184401251?pgno=2`, 2000. Online; retrieved 01.11.2019.

[Po69] Postel, Hans Joachim: Die Kölner Phonetik. Ein Verfahren zur Identifizierung von Personennamen auf der Grundlage der Gestaltanalyse. In: IBM-Nachrichten, 19. Jahrgang. pp. 925–931, 1969.

[SPS15] Schürer, K.; Penkova, T.; Shi, Y: Standardising and coding birthplace strings and occupational titles in the British censuses of 1851 to 1911. Historical Methods, pp. 195–213, 2015.

[Wi90] Winkler, W. E.: String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. In: Proceedings of the Section on Survey Research Methods. American Statistical Association. p. 354–359, 1990.