10 March 2023

English

United Nations Group of Experts on Geographical Names
2023 session
New York, 1 – 5 May 2023
Item 14 of the provisional agenda [*]
Geographical names data management

The gazetteers.net web application: a tool for harvesting digital gazetteers

Submitted by Germany[**]

Summary

Having access to and the availability of many types of digital gazetteers reduces the effort involved in conducting place-name research. Some digital gazetteers cover the whole world and diverse languages. However, the coverage of separate regions varies. In addition, global gazetteers marginally cover small regional languages, which in turn become the focus of small initiatives, some of which are managed by one person. Moreover, digital gazetteers usually do not reflect administrative changes, such as incorporated towns, which are often not represented. Since there is no standard definition of place as a geographical unit (human settlement), the scope of places that are mentioned in the individual gazetteers include individual farms, mills and municipalities.

The sheer number, different geographical coverage and metadata schemes of digital gazetteers make it difficult to compare existing gazetteer entries systematically and to use existing data in other applications. At the same time, current digital gazetteers show how geographical orders of knowledge are transformed from analogue structures (for example, printed indexes) into digital structures. The gazetteer research project in Germany, undertaken by the Herder Institute for Historical Research on East Central Europe (Marburg), the Institute for Regional Geography (Leipzig) and the Justus Liebig University Giessen, has developed a publicly operational web application, gazetteers.net. The objectives of the application are to support users working with different digital gazetteers and to help them explore content and metadata structure of the gazetteers.

The gazetteers.net web application enables users to search several place name-related databases simultaneously in a unified manner and to view and compare data from different gazetteers. In addition, the application supports the identification of items in different databases that refer to the same geographical entity, regardless of the definition of geographical place in the individual gazetteers or its administrative status. By linking corresponding items across gazetteers, the application facilitates data aggregation and

[*] GEGN.2/2023/1

[**] Prepared by Ihor Doroshenko, Christian Lotz, Dariusz Gierczak and Francis Harvey (Germany), Herder Institute for Historical Research on East Central Europe and Institute for Regional Geography.

comparison. In addition to the major and well-known web gazetteers, the official gazetteers and some small local gazetteers for a selected country (Poland) have been connected in order to be able to cover regional languages and historical names. A comparison of these specific and general gazetteers has also facilitated, among other things, the identification of differences regarding languages, spelling and administrative changes throughout history.

The project team has examined existing digital gazetteers for their structure (semantics, description of metadata) and content (reliability of assignment between place names and coordinates). The project team has also discussed geographical discourses inherent in existing gazetteers and examined strategies to reveal specific power-knowledge relationships within existing gazetteers. Having examined the results of this testing, project participants revised and refined the metadata structure and web application interface.

The recent version of the harvesting tool was launched online after a positive evaluation by the expert communities. Despite the current regional focus of the project, searches can also be conducted at the global level. Current work on the tool is aimed at finding a way to incorporate more gazetteers, for example, of other countries or regions, without sacrificing clarity and responsiveness. Since the application is designed to support searches in the existing gazetteers, the quality of the results depends directly on the quality of each connected source.

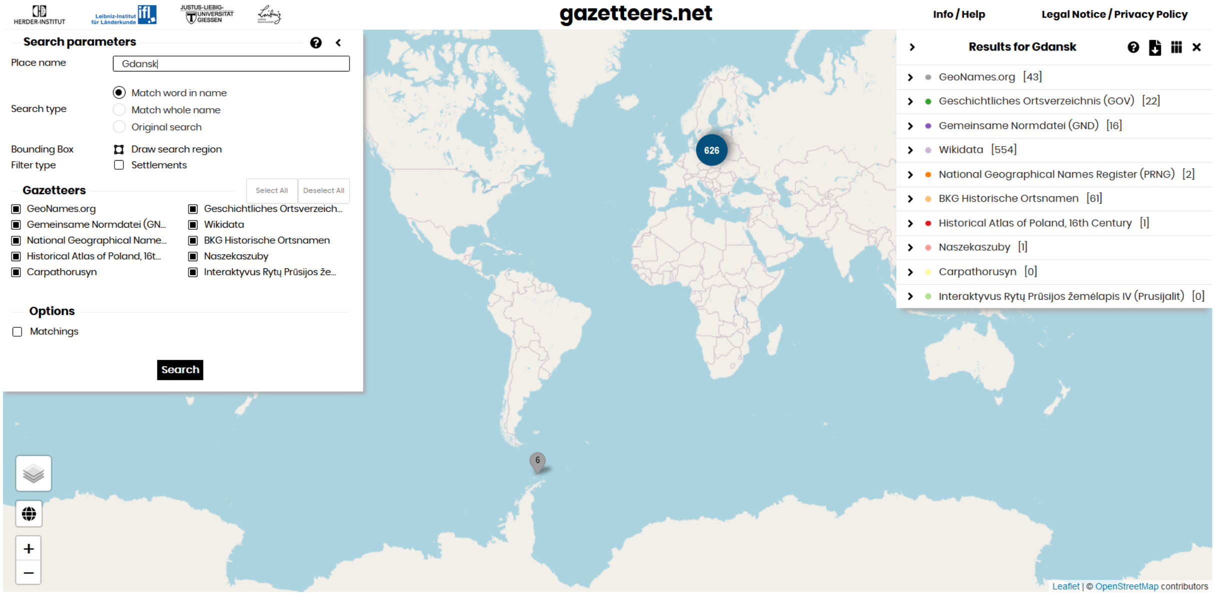
Introduction

Management of data related to geographical names is a challenge for many branches of the humanities and sciences, as well as politics and administration. The reasons for these challenges are manifold, since geographical names may have changed over time. Furthermore, different ethnic or social groups may have different names for the same geographical object, which can result in political disputes about the presumably 'correct' name and changes in official names.

Throughout history, various institutions, citizen science initiatives, and individual actors have collected and recorded geographical names to settle the challenges or influence rivalries about place names. Once, a geographical name has found its way into a gazetteer, the gazetteer entry becomes a more or less fixed point of reference. In other words: gazetteers create specific orders of geographical knowledge. For about two decades, an increasing number of digital gazetteers have been overcoming traditional (i.e. printed) indices of geographical names. The transformations from analogue to digital gazetteers produce an ambivalent situation: On the one hand, available digital gazetteers, for example, Geonames, facilitate the search for geographical names. It is an enormous support for research in the sciences and humanities and for administrative purposes to search for geographical names digitally. On the other hand, the content and the metadata structure create new orders of geographical knowledge, for example, by defining a name as an 'official' name or fixing the spelling of a name in favour of one dialect over a neighbouring dialect.

It is against that background that the Herder Institute for Historical Research on East Central Europe (HI), the Institute for Regional Geography (IfL) and the Justus Liebig University Giessen (JLU) set out to investigate these processes of transformation of gazetteers and develop a tool to systematically search in various digital gazetteers as well as to compare metadata structures and content of these gazetteers.

HI, IfL and JLU are academic institutions, and the history and geography of Central and Eastern Europe belong to their main areas of research. All three institutions receive financial support from the German government, and the Leibniz Association supported the Gazetteers project. Neither HI, IfL, nor JLU maintains any official gazetteers. Instead, we perceive place names and gazetteers (printed and digital) as research sources. At the same time, we hope that the application Gazetteers.net will be helpful not only for researchers, librarians and archivists but for all kinds of data curators, data management officials, and citizen sciences initiatives dealing with place names.


HERDER-INSTITUT
Leibniz-Institut für Länderkunde ifl
JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN
gazetteers.net
Info / Help
Legal Notice / Privacy Policy
626
6
+
−
Leaflet | © OpenStreetMap contributors

Search parameters

Place name: Gdansk

Search type: ● Match word in name ○ Match whole name ○ Original search

Bounding Box: Draw search region

Filter type: □ Settlements

Gazetteers Select All Deselect All

■ GeoNames.org ■ Geschichtliches Ortsverzeich...
■ Gemeinsame Normdatei (GN... ■ Wikidata
■ National Geographical Name... ■ BKG Historische Ortsnamen
■ Historical Atlas of Poland, 16t... ■ Naszekaszuby
■ Carpathorusyn ■ Interaktyvus Rytų Prūsijos že...

Options

□ Matchings

Search

Results for Gdansk

- GeoNames.org [43]
- Geschichtliches Ortsverzeichnis (GOV) [22]
- Gemeinsame Normdatei (GND) [16]
- Wikidata [554]
- National Geographical Names Register (PRNG) [2]
- BKG Historische Ortsnamen [61]
- Historical Atlas of Poland, 16th Century [1]
- Naszekaszuby [1]
- Carpathorusyn [0]
- Interaktyvus Rytų Prūsijos žemėlapis IV (Prusijalit) [0]

Figure 1: Gazetteers.net application

The main interest of our project was to compare data and metadata structure, which are provided in gazetteers for a certain region. We selected Poland as the case study, as there have been many changes in borders and migration processes since early modern times. In our application, we include data from three different types of gazetteers: 1) state official gazetteers, such as Gemeinsame Normdatei (GND, Integrated Authority File, Germany) and Państwowy Rejestr Nazw Geograficznych (PRNG, National Register of Geographical Names, Poland); 2) gazetteers which are created by citizen science initiatives and which are widely used in Central and Eastern Europe, i.e. Wikidata, Geonames, and Geschichtliches Ortsverzeichnis (The Historic Gazetteer, GOV); and 3) gazetteers which had their origin in research projects, such as Kaszëbsczé miestné muiona (Kashubian place names), Carpathorusyn (The Lemko Village Resource Guide) and Interaktyvus Rytų Prūsijos žemėlapis V (Interactive map of Eastern Prussia, pt. V).

Search, Representation, and Analysis

There are many ways how the Gazetteers.net application can explore data from existing databases. Our project focuses on comparing place names as the lowest common denominator among existing databases. Data exploration consists of three basic parts – search, representation, and analysis (see figure 2).

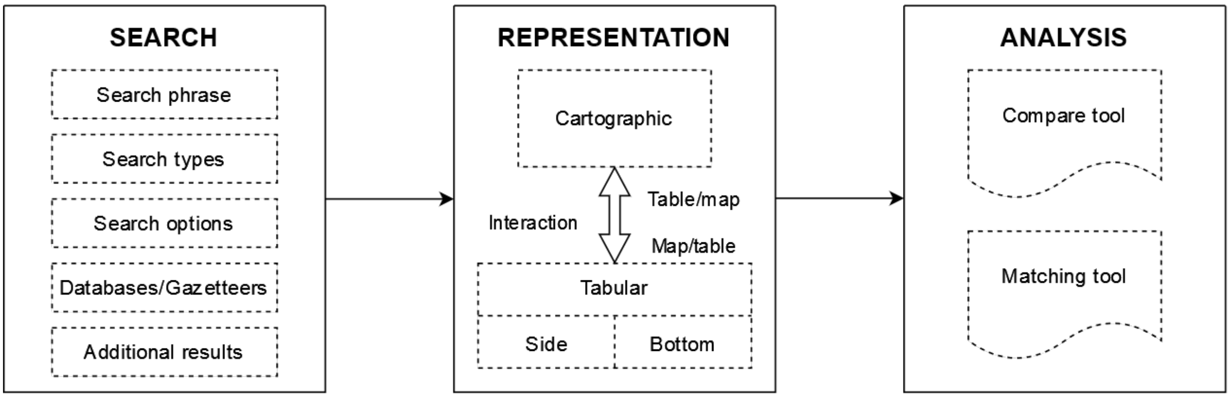


Figure 2: Workflow of the Gazetteers.net application

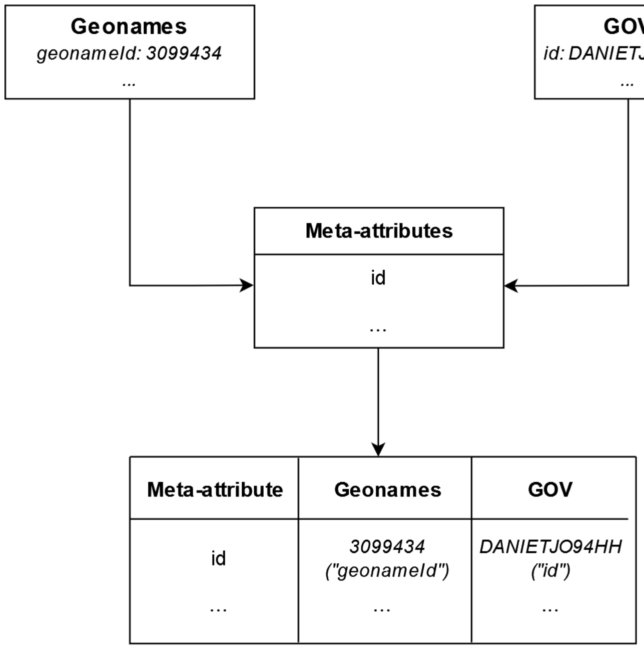
In the search part, the user inputs a place name and the search parameters, which include search type and some restriction options to narrow results to find the requested place name. The application supports the identification of corresponding items in different databases. It identifies possible “same as”-relationships for every item in the search result set. The name search using Latin-based characters does not require the consideration of grammatical case- and diacritical marks.

Some gazetteers contain “same as”-references to other gazetteers, which allows identification during the life query. Furthermore, the possible 'same as'-relationships are already included in a separate project database, as determining them live during a user search would take too much time, especially for large result sets.

Search type defines how it will be searched lexically. There are three search type options in the application – “Match word in name”, “Match whole name”, and “Original search”. The first type of search defines whether the search phrase only needs to match a word in the name (search word “Gdańsk” - or, as mentioned above, the notation „Gdansk“ is equal - also in the results, there are those entities which contain this name such as “Gdańsk Brętowo”). The second search type indicates whether it must match the complete name of an entity (search word “Gdańsk”, in the results, there are only those entities which have the name “Gdańsk”). The third search type uses the original settings of the databases. Other search parameters include restriction options that are either feature-based (via selecting the option “only settlement” that enables search for only settlement place names), or geographical (via defining the bounding box on the map where it should be searched). After that, the user selects databases to search in. Finally, the option to enable matchings should be selected, if the user wants to get matchings to other databases in the results as an additional attribute.

After the user inputs the search phrase and all parameters, he starts the search. The representation of results has two forms – cartographical and tabular. The cartographic presentation dominates the user interface. The tabular presentation can be shown inside or bottom view modes. The result table is split into sub-tables, each representing gazetteer-specific results. However, as the sub-tables are modular, they provide the same possibilities of interaction with the map, exporting, sorting, filtering the data, etc.

Several concepts are running in the application's background to enable analysis functionalities. One of them is the concept of comparing entities. It describes principles and methods to compare entities of different data sources (gazetteers), quality, formats, and data types. For this aim, a uniform meta-data schema was created. The entities are compared according to this schema based on their attributes. It should explore and find similar attributes (for example, “location”) in the entities selected to compare. Notably, the entities are compared based only on those attributes available across the gazetteers used in the application. It includes such meta-attributes as “id”, “name”, “variant names”, “position”, “type”, and “link” (see figure 3). The most important criterion for the selection was the commonality of the attribute over all of the included gazetteers in one or the other form. Even if such attributes as “type” or “position” in different gazetteers can have very different structures, they should still be matched with the meta-attribute “type” or “position”. If the meta-attribute is not found, the program reacts dynamically and shows space for the entity. This schema is also extensible. In other words, in future developments, more attributes can be added to it.

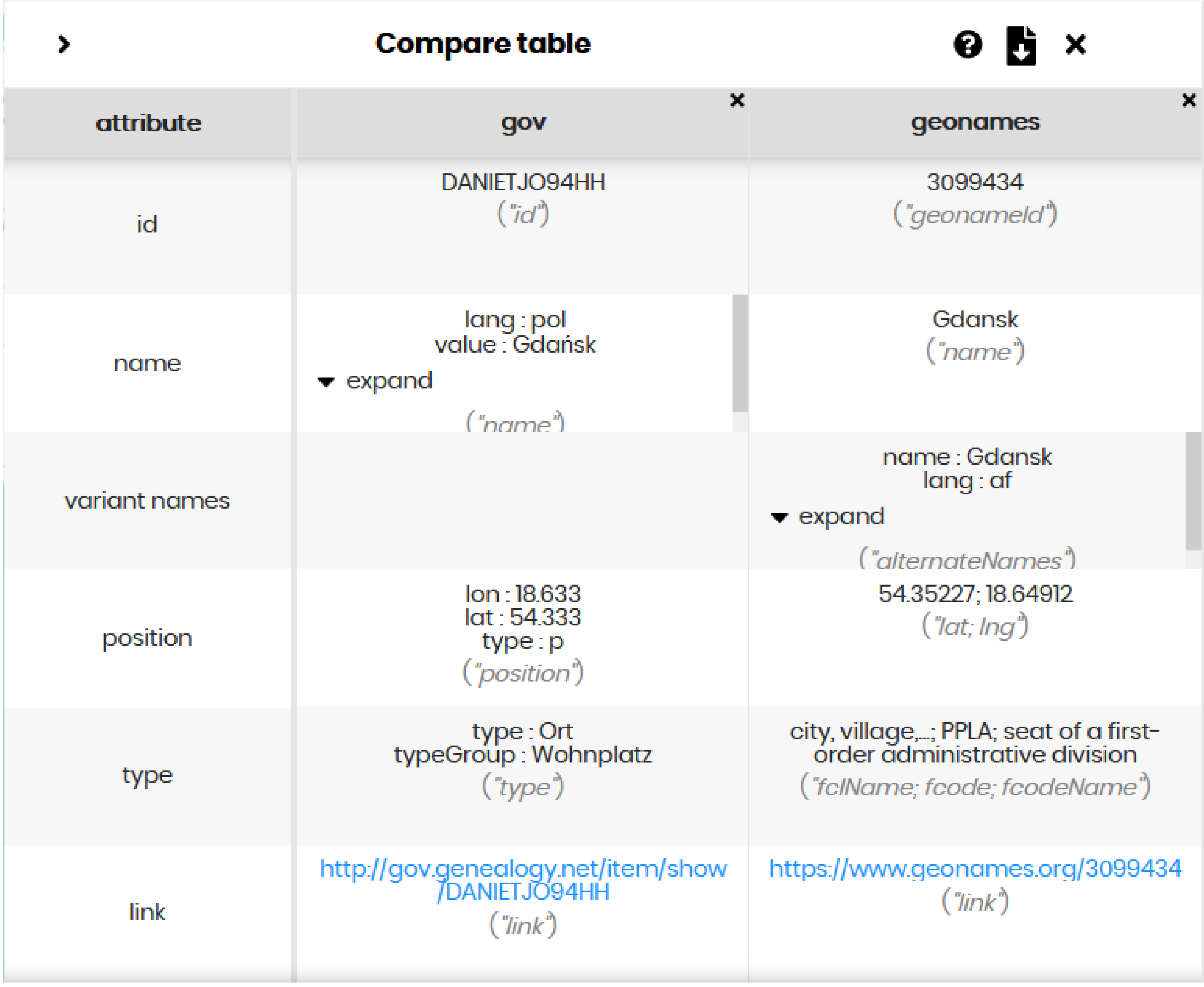


Meta-attribute	Geonames	GOV
id ...	3099434 ("geonameId") ...	DANIETJO94HH ("id") ...

Figure 3: Example of a meta-attribute schema for id-attribute for the compare tool

Furthermore, it should preserve original attribute names, as some gazetteers can contain such information as longitude and latitude in attributes called "location" or "coordinates". The attribute names are, in this case, different, though, but they contain the same information type, namely, spatial information. Original attribute names matched to the meta-attribute names are shown in the value cell of the entity (see figure 4). This way, one can compare attribute names to understand how the gazetteers handled the "containers" for their information.

However, this approach provides some important limitations, such as information loss. For example, the attributes that provide information about the administrative hierarchies (which indeed can be important for the comparison) are unavailable across all the application's gazetteers. Only some gazetteers, such as Geonames, GOV, and GND, provide this information. But, even provided, this information is difficult to compare. GOV provides only general administrative information, both modern and historical, while Geonames does it very detailedly, but only for the present time. As this attribute cannot be equally compared, it is neglected and not considered in the compare tool of the application. However, the application allows the export of the entities selected to compare in different formats like CSV, JSON, and GeoJSON for further processing.



Compare table

attribute	gov	geonames
id	DANIETJO94HH ("id")	3099434 ("geonameId")
name	lang : pol value : Gdańsk ▾ expand ("name")	Gdansk ("name")
variant names		name : Gdansk lang : af ▾ expand ("alternateNames")
position	lon : 18.633 lat : 54.333 type : p ("position")	54.35227; 18.64912 ("lat; lng")
type	type : Ort typeGroup : Wohnplatz ("type")	city, village,...; PPLA; seat of a first-order administrative division ("fclName; fcode; fcodeName")
link	http://gov.genealogy.net/item/show/DANIETJO94HH ("link")	https://www.geonames.org/3099434 ("link")

Figure 4: Compare tool in the Gazetteers.net

Another important feature of the application is the matching algorithm. Using the term "matching" signals a flexible approach to comparing place names using algorithmic enrichments, such as Levenshtein distance[1]. Corresponding entities can also be identified by comparing their attribute values like names and coordinates, especially their combination. Furthermore, similarity measures can be computed for two names (using the Levenshtein distance) or two coordinates (calculating the geographic distance), when such information sets are available in the corresponding data sets. In many cases, the data situation does not allow a clear assignment based on such matching strategies because of the place name variations, missing coordinates, differences in the particular gazetteers or just the absence of a record in a database. To reduce ambiguous matchings, the Gazetteers.net web application compares the single entities based on normalized data (e.g. optionally removing diacritics and name affixes). Additional information, like the entity type, is in use here. Like the "reference table", the "matchings" component was built offline and is stored in the application's database. As the matchings lookup requires additional search operations for each entity in each result set, and in sum, this can be time-consuming,

[1] Kessler, B. (1995). Computational dialectology in Irish Gaelic. In S. P. Abney / E. W. Hinrichs (eds.), Proceedings of the seventh conference on European chapter of the Association for Computational Linguistics. Morristown: Association for Computational Linguistics, p. 60, https://doi.org/10.3115/976973.976983; Angelis, G. de, Jessner, U., & Kresic, M. (eds.) (2017). Crosslinguistic influence and crosslinguistic interaction in multilingual language learning. London: Bloomsbury Academic; Schepens, J. J. (2008). Distributions of cognates in Europe based on the Levenshtein distance. Radboud University Nijmegen, retrieved from https://theses.ubn.ru.nl/bitstream/handle/123456789/47/Schepens,%20J.%20BaThesis.pdf?sequence=1.

this feature can be turned on and off. The complete system of identifying possible "same as"-relations in the application is shown in figure 5.

Although local storage of some contents significantly improves the application's responsiveness, it has limitations, such as local storage resources and the availability of long-term hosting.

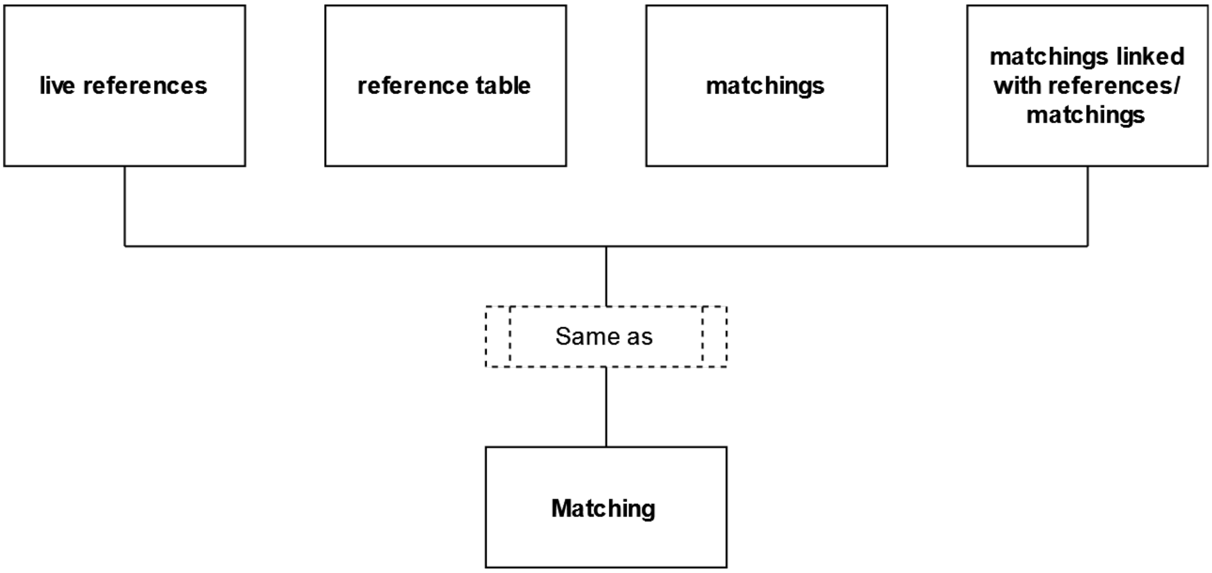


Figure 5. System of identifying possible "same as"-relations in the Gazetteers.net application

Conclusion

Databases which collect place names still develop and find many important uses. This applies particularly to citizen science initiatives like Wikidata or university-based research projects like the World Historical Gazetteer. Our application, Gazetteers.net is open source (for details, see the help section at Gazetteers.net), as we want to improve the capability to compare differences in existing databases. These differences, for example: what is the 'preferred' place name, may have enormous political, social or ethical implications. We hope that our application Gazetteers.net helps to explore the content and metadata structure of existing gazetteers and opens up perspectives for further research questions.

Acknowledgements

The creation of this paper was supported by Eric Losang and Marc Friede. The project team is grateful for having been able to cooperate with Jesper Zedlitz (Geschichtliches Ortsverzeichnis / The Historic Gazetteer), Esther Scheven (Deutsche Nationalbibliothek / German National Library), Wiesława Duży, Katarzyna Słomska-Przech, Tomasz Panecki and Marek Słoń (Polska Akademia Nauk / Polish Academy of Science) as well as the Bundesamt für Kartographie und Geodäsie (Federal Agency for Cartography and Geodesy).

Points for discussion

The Group of Experts is invited to:

(a) Take note of the report and progress made by the project;

(b) Express its views on the harvesting tool.