Data Integration in Support of Gazetteer Development

Bruno Magalhães

Gazetteers provide information on geographical places and their respective spatial location and relationships, with several projects over the years developing different frameworks to better support the management of this type of data. Furthermore, the creation of a gazetteer database also requires the implementation of methods capable of integrating data from external sources using different data models as well as the delineation of spatial and temporal regions. However, there still isn't a dedicated software framework capable of handling both the base data management of a gazetteer and the challenges of integration and region delineation. This research thesis focus on the creation of an all encompassing gazetteer that provides a functioning database working alongside the necessary integration and delineation tools. To achieve this, a framework based upon previous gazetteer projects will be used as a basis upon which the integration and delineation tools will be implemented. The final gazetteer will then be made available through a web application interface which provided a multitude of ways to interact with the gazetteer's data including data export and schema interaction. Its database contained a large amount of data with a focus on the early colonial Mexican period whose accuracy was evaluated based on models for determining gazetteer quality and the geographic distributions of the places within the database. This evaluation largely confirmed the accuracy and consistency of the data available in the gazetteer also by comparing it with other alternative external sources. The main tools utilized during the data integration process for this data were a duplicate detection tool and polygon data generation tool. Both of these tools were evaluated in scenarios specific to each of them with the former being able to filter out a good number of duplicates and the latter accurately generating spatial regions for places from reference points.

Keywords: Historical Gazetteer, Data Integration, Geometry Generation, Visual Interface, Data Export

1. Introduction

A gazetteer is a database of geographical places which stores information about their designations, type, location and many other descriptive features (Berman et al. 2011). The main purpose of a gazetteer framework is to provide answers to questions about existing places (e.g. *Where is Paris?* or *How many designations exist for New York City?*) including how the collective of places in database relate to each other either through a hierarchy or a more nuanced relationship type. This requires a way to effectively represent the concept of a place within the format of a digital gazetteer (Purves et al. 2019). To further support this objective gazetteers are often accompanied by virtual tools such as digital maps for representing the place's location (Acheson et al. 2017) or timelines in the cases where temporal information is present in their database.

Despite this, the development of gazetteers presents many challenges (Berman 2016) particularly in the gathering and processing of information on the relevant places. The lack of a unified source for place data leads to the use of many different sources often with conflicting or duplicate data. Using other gazetteer as sources also presents problems since they may use a completely different type of data organization which may hinder the integration process.

Taking this scenario into account, this research projects attempted to create a gazetteer capable of integrating data from multiple sources and present it in a universal structure for easier future use. The gazetteer also includes temporal data for places with historical relevance. Previous gazetteer projects were used as an initial foundation to facilitate the creation of this gazetteer framework. The contributions of these related gazetteer projects will be detailed further in Section 2 of this article.

Therefore, the first objective of this project consisted on creating a functional historical gazetteer capable of importing and integrating data from multiple sources. The next main feature is to present said data on a digital interface that allows any user to explore and interact with the data in the gazetteer while also providing the option to export the place data in a universal format of choice (i.e., Linked Places Format, CSV or Shapefile). The export feature greatly increases the potential of the gazetteer as a universal source for place data effectively bypassing the challenges of the initial import and integration process that comes with using traditional sources.

The rest of this article presents the aforementioned related works which helped setup this project alongside the architecture and functionalities of the historical gazetteer created during this project (Section 3). Finally, a conclusion and possible future work is presented at the end of the article.

2. Related Work

While the concept of providing a historical gazetteer framework with data integration and export capabilities in a global format is novel, the more basic aspects such as the database schema organization, the data formats and the interface used take ideas and formats from other already working gazetteer projects.

Starting of, the entire schema of this gazetteer project is based around the proposed schema from the Alexandria Digital Library (ADL; Hill et al. (1999)) project. ADL is one of the first digital gazetteer projects and one of the pioneers in the transition from physical to digital gazetteers. To achieve this goal the ADL project created a definition

of *place* that could be used in the context of a digital database. This definition primarily divides a *place* into 3 main elements: placename, placetype and geographical footprint. Furthermore, a temporal data field was added to take into account possible temporal information of the place in question. The resulting schema provides the necessary framework to effectively describe the gazetteer entries themselves as well as the relations and relevant metadata between them. To make use of the placetype and temporal fields it was necessary to import metadata for relevant placetypes and time intervals.

The previous described ADL format requires a hierarchical format to make use of its classification features. To this end it was necessary to find a source with a usable classification scheme to populate the ADL format chosen. Among the many existing formats the most interesting project for this was the Getty Thesaurus of Geographical Names (TGN; Harpring (2010)). The main contribution for this project comes from the TGN's focus on hierarchical and internal relations between places that it stores. The TGN attributes a type to each place which can be one of the following categories: administrative entry, physical feature or metadata types. There also exist relation types with the resulting hierarchy between places allowing the user to trace back a place all the way to the largest entity within the gazetteer database. This type of classification organization was imported into our gazetteer framework essentially meaning it shares the same placetype hierarchy of the TGN.

The next relevant work was the World Historical Gazetteer (WHG; Manning (2015)) project. Although still under development, the WHG's data export features is the main contribution used in this project. Besides its gazetteer framework, the WHG makes use of the Linked Places Format (LPF) as the main data format within its gazetteer database. This data format proves to be very useful as a global format capable of being used by multiple different gazetteers. In particular, the LPF allows the search and linking of terms across different gazetteers from its stored sources alongside information gathering to differentiate separate places from multiple gazetteers. These global integration advantages prove to be an important functionality in an environment where many gazetteers make use of distinct data formats. Besides its potential as a global gazetteer data format, the LPF also includes naming, temporal, location and hierarchical information on the place in question keeping the capabilities of being a format for individual place information.

Other gazetteer projects, such as the GB1900 (Southall et al. 2017), project focus instead on using a large number of volunteers to gather large amounts of data. The main goal of GB1900 is the creation of a highly detailed historical gazetteer of Great Britain, to be made available to the public in universal data formats. This crowd funded approach will allow the GB1900 project to achieve an unprecedented level of detail regarding the data present in its database.

The last relevant project to highlight is the Who's on First? [1] project which itself is also another large coverage gazetteer. However, the most important element from this project is its visual interface Spelunker. This tool provides a simple interface that allows the user to explore the places and data stored in the gazetteer. This includes all the information belonging to specific places such as possible denominations, the geometry with the corresponding position in the world map, their type and other related places. The user interface of our project was adapted from the Who's on First? user interface in order to better fit our needs.

[1] `http://www.whosonfirst.org/`

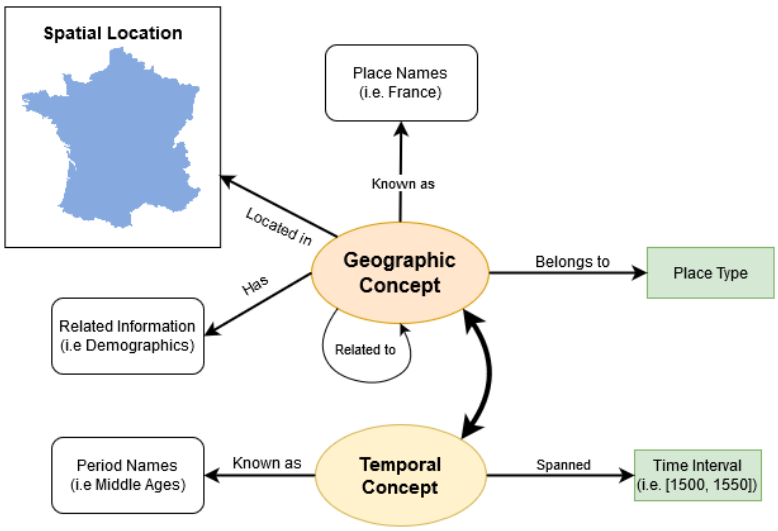

Figure 1.: Concept and components of a place in the gazetteer.

3. A Toolset Supporting Gazetteer Development

The structure and development of the gazetteer can be divided into the following main parts: (i) the handling of data within the gazetteer's database, (ii) the user interface aspect, and (iii) the delineation of vague regions for places with unknown limits.

The core of our gazetteer makes use of contributions from some of the aforementioned explored gazetteer projects. The database schema makes use of the ADL schema format which provides a stable base to handle place data in its geographical, type and temporal dimensions. Figure 1 showcases the concept and components of a place in the database of the gazetteer which forms the logical base for the data schema used.

The data used to populate the gazetteer can be divided into 2 groups: classification data and place data. Regarding classification data, in order to make use of the adopted ADL schema it is necessary to gather terms for type, temporal and source categorization. Like previously stated in Section 2, the TGN project offers a useful list of terms to categorize different places. Furthermore, in the context of associating time periods with common denominations (e.g. *world war 2* being the year interval of 1939 to 1945) the PeriodO [1] project contains a vast database of these term to time period associations.

The next type of data used refers to the actual place data (e.g. names, location and time period) which populates the vast majority of the gazetteer database. The first source for place data is the GeoNames project which itself is a database of geographical data acquired from multiple other sources. The data in GeoNames is made available through a variety of web services that allow access to the project's data in a text format and said data is daily updated according with changes to the GeoNames project.

However, the main bulk of place data comes from the Digging into Early Colonial Mexico (DECM) [2] project. The DECM project develops a digital approach to explore data from the early colonial Mexico period. The data was assembled from a collection of Spanish official documents from the era available in the corpus *Relaciones Geográficas de la Nueva España*.

[1] http://perio.do/
[2] http://www.lancaster.ac.uk/digging-ecm/

3.1. *A Web Application for Managing and Exploring Gazetter Data*

The web application acts as the frontend interface for the entire historical gazetteer framework which includes a basic searching of place at the top of the page and tools for the direct interaction with the database schema and records. Therefore, the web application can logically be divided according to these 2 levels of access and their corresponding features. Taking this division into account both parts will be described separately in the remainder of this subsection for easier understanding.

The first section of the web application belongs to the basic access level of the web application. The main focus of this component is to allow users to easily search for places stored in the gazetteer's database. The user can then select places from the list that results from the search query which opens up the individual information page for that specific place. This page contains more in depth data on that specific place such as its alternative names, placetype, coordinates and all other related entries in the gazetteer.

Furthermore, this component also allows to export data in various formats such as the previously detailed LPF format from the World Historical Gazetteer project. The data export functionality stretches from a single selected place to a group of places that result from a search query or even the entire database if requested. This functionality allows the data from the gazetteer to be easily used in other projects of the same type bypassing the process of explicitly importing and integrating the data.

Additionally, upon logging into the application the user gains access to the advanced level of the application which provides a greater array of data management options. In this level the user is able to interact directly with the database and perform changes to the schema or inserting new entries into the gazetteer.

3.2. *Data Integration and Duplicate Detection*

Another main component of this project is the process of data import and integration into the gazetteer database. The first step of this process is to gather the data from the sources chosen to populate the gazetteer which can be divided into 2 parts. An initial import of metadata for placetypes and temporal periods which can be used to classify the places of the gazetteer. The sources used for this initial import are the TGN and the PeriodO [1] projects database, respectively. The second part represents the main bulk of data by importing primarily data from Geonames (Wick and Vatant 2012) and the DECM [2] project.

The information related to each individual place is still lacking in the inter relationships between places and their valid temporal period aspects. This is mostly the result of the sources utilized simply not detailing both these components of a place. More relationships between places were determined using the geographic footprints of each place and calculated how it related to all other places in the gazetteer. This process primarily consisted on analysing all places with each other to determine their spatial relationship such as adjacency, overlap and if places are contained within other places. Additionally, the temporal information of each place was determined by analysing historical documents and identifying which places in the gazetteer were mentioned in those historical documents.

Finally, after the processing of both relational and temporal place data the last step involves the detection and removal of duplicates that still remain in the database. First,

[1] `http://perio.do/`
[2] `http://dh2018.adho.org/en/a-deep-gazetteer-of-time-periods/`

all places with defined locations were analysed among each other to determine which pairs of places had a similar location with the threshold being a distance of less than 500 meters between both places exact location. In the case of more complex spatial footprints such as polygons both regions were analyzed and in case they are considered identical they were also marked as a possible duplicates. Afterwards, the names of the possible duplicate places were compared through a string similarity algorithm (i.e. Jaro-Wrinkler) with a threshold similarity of 0.9 (meaning above this similarity it was considered a duplicate pair).

In the case of a positive match in both of these conditions, the pair of places were considerate duplicates and were marked as such with a new relation of equivalency between said places. This leaves the possibility for future updates where the duplicate pairs can have one of the places be removed, both places merged into a single entity or only one of them is considered as the main place and visible in the web application interface.

In order to evaluate the performance of this tool it was necessary to find a scenario with a number of duplicates and then run the duplicate detector in said environment. As such, the best way to evaluate the accuracy of this tool is to test it on a large sample of duplicates. For the purpose of this evaluation a dataset from the Geonames Wick and Vatant (2012) project containing a large number of placenames pairs with the corresponding classification annotations was used. However, it was still deemed important to check the tool's performance on the gazetteer itself so a sample of 50 duplicates was randomly retrieved and manually evaluated to give an insight if its performance in the data integration process.

3.3. *Inferring Spatial Footprints for Place Records*

Finally, the last main component of this research project was the polygon data generation tool created to handle places with missing or vague spatial information. During the development of the gazetteer it was noted that some places still lacked proper information regarding their location and type of geographical geometry. However, some of these places with vague geographical information still had some points of reference such as boundary boxes and other related points which could be used as reference to create a more well defined spatial footprint.

A script was developed to generate polygons corresponding to the actual area of these vague places. The script receives a list of reference points as an input and creates a Voronoi diagram (Voronoi 1908). However, from initial tests it was noted that this approach on its own was lackluster so a new variable was added to the diagram creation process which was weights based on population values from the Gridded Population of the World [1]. Essentially reference points with higher population weight were given a higher value for the distance based algorithm for attributing units of area to a polygon in the diagram.

The resulting polygons were then extracted and attributed to each relevant place that contributed to its generation or in other words the places whose reference points were contained in a polygon were attributed with said polygon. This component assures that places with missing information can also have a complete representation in the web application user interface and also allows for further contextualization in the overall gazetteer hierarchy from its spatial relationships.

[1] https://sedac.ciesin.columbia.edu/data/collection/gpw-v4

Evaluation of this tool revolved around attempting to recreate regions already present in the gazetteer database. In particular, reference points were retrieved from a set of places with polygon representations of the same placetype and then utilized as an input to the polygon generation tool. Afterwards, the generated set of polygons were compared with the original ones from the database in order to determine how faithful the recreation attempt was. This approach allowed for an easy way to evaluate the tool while also allowing the option highlight its strengths and weaknesses through the use of specific scenarios catered to each one.

4. Results and Analysis

4.1. *A Case Study with Early Colonial Mexico Data*

The process of evaluating a gazetteer is not straight forward since there isn't an exact expected result prior to its creation or an universal evaluation metric. It was necessary to find a model with a method capable of determining the quality of the final state of this gazetteer. The chosen evaluation method was the one described in the Acheson et al. (2017) research project which focused on analyzing and comparing global gazetteer projects such as the GeoNames and TGN, also used as sources for our historical gazetteer, by classifying them according to a predetermined set of criteria and through an analysis of the distribution of places in the respective gazetteer.

At the moment of writing, the gazetteer contained 47305 features (individual places) stored in the database. In terms of place names or designations, the gazetteer contained 56118 place designations with only 41205 being actual distinct names.

Place distribution and locations can also help determine the overall accuracy of a gazetteer. The database of our gazetteer contains a total of 66 distinct placetypes. By selecting places by their type and analysing their distribution allow us to check if the locations are in line with their corresponding type. For example, places with the type *harbour* should be near the coastline or at least near large bodies of water.

Furthermore, the integration process also created a network of relations between all the places within the gazetteer which helps contextualize each individual entity into the larger scope of the entire gazetteer. Currently, the gazetteer contains a total of 228914 relations among the places stored in the database with an average of 26 relations per place and the maximum number of relations for a single place going all the way up to 7294. Each relation has a type specific to it and the percentage division of all types of relations available in the gazetteer was also analysed. It is important to highlight that the majority of the relations correspond to places contained within other places (i.e. *Member is* and *Member of*) and the number of duplicate relations belonging to the type *Possibly related to*. However, the total number of duplicates is only half of the relations of duplicate since each place has the relation of duplicate with its other identical place.

Overall, the gazetteer in its final state proves to be an accurate representation of the data provided by the sources used, in particular the DECM project. These results are directly related to the quality of the integration tools used which in turn makes it relevant to also evaluate the major tools used during that process in this section.

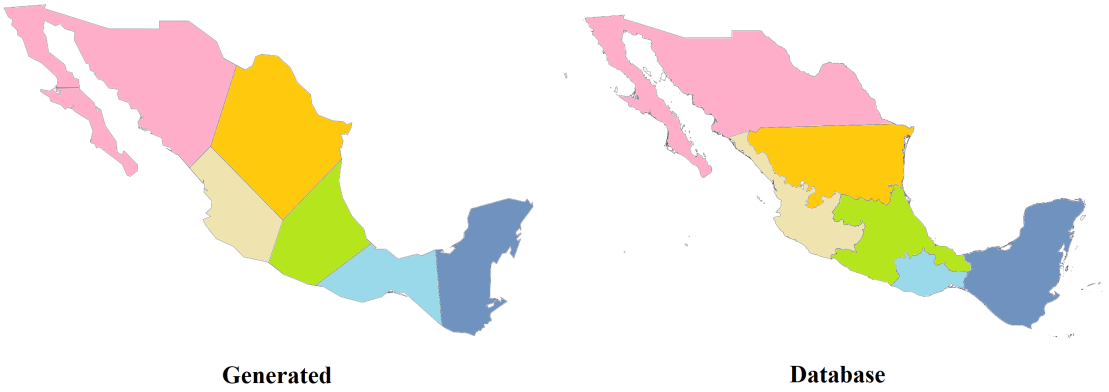

Figure 2.: Side by side comparison of Mexican Geographical Regions.

4.2. *Results for Duplicate Detection*

The results of running the duplicate tool mechanism across the GeoNames dataset are represented in Table **??**. The duplicate detector managed to detect 341122 duplicates correctly and correctly labeled another 824298 as not duplicates for a total of 70.5% accuracy. A more detailed look shows a sensitivity value of 93.77% which demonstrates a very low number of distinct places being wrongly considered as duplicates. Meanwhile, the specificity metric sits at the much lower value of 44.07% and is the main factor bringing down the accuracy of the duplicate detection tool.

Consequentially, the results of this test point towards a bias for limiting the number of pairs being wrongly considered as a duplicate at the expense of leaving out some real duplicates as false negatives. Whilst the ideal scenario would be the correct classification of both positive and negative duplicates, having a preference for avoiding wrong duplicated classifications at the expense of some duplicates not being detected is a better option than the opposite. Considering that duplicate entries might be merged or one removed in the future this could result in a large number of non-duplicate places being wrongly removed which essential means the loss of an important distinct place.

Likewise, the alternative test of randomly selecting 50 duplicates from the gazetteer database provided similar results. The results achieved an accuracy of 90% (45 out of 50) which is in line with the test on the considerably larger GeoNames dataset. In particular, the high sensitivity of that previous test is reflected on the large number of correctly determined duplicates of the test in the random 50 duplicates from the gazetteer database.

From these results, it is possible to conclude that the duplicate detection tool had a good performance in detecting duplicates in the gazetteer data integration process. However, the low specificity values raises some questions regarding possible duplicates being missed due to the tool's more conservative approach on marking one as such.

4.3. *Results for Inference of Spatial Footprints*

The methodology for this evaluation focused on comparing generated polygons with their original representations in the gazetteer database. Figure 2 showcases an example of these scenarios that being the attempt to generate Mexico geographical regions by the data generation polygon.

Both representations share a similar division of polygon areas which means the polygon generator tool managed to recreate the original regions with an acceptable degree of accuracy. There some shortcomings in the recreation like the lower accuracy on recreating the irregular shape of some regions, most notably the central green polygon. Most scenarios utilized follow this general trend of the polygon generation tool being able to recreate to a large degree the original division and distribution of polygons.

Other weakness were also noted in some specific cases of trying to recreate regions with a set of conditions that presented problems. In particular, the tool presented issues in dealing with neighbouring regions sharing vastly different sizes since the tool tended to try and balance both sizes resulting in the smaller region ending up much larger then what it was supposed to be. Furthermore, another scenario consisted on providing only a portion of the total states of Mexico resulting in some of states being much larger than supposed by taking over area from the missing states not used as input.

Despite these particular issues, the overall performance of the tool is more inline with the example scenario provided in Figure 2 with the recreated regions following the original ones closely. These results provide assurance that the number of places with generated spatial footprints in the database have a high probability of being inline with the actual place's geographical presence.

5. Conclusion and Future Work

At the start of this project we established the goal of creating a gazetteer framework capable of handling data management and integration challenges. Analysing other gazetteer projects (e.g. ADL and WHG) allowed us to base our data formats, schema and gazetteer functionalities according to these projects contributions. At the time of writing, the gazetteer, alongside its respective web application interface, contains all the functionalities it set out to achieve which in turn is one of the major positive aspect of this entire project.

Additionally, parallel to the creation of the gazetteer database and interface, tools for the data integration process were developed with the major ones being the duplicate detector tool and the polygon data generator. The former provided overall positive results when handling duplicates with similar locations and primary designations. However, its results also showcased the conservative behavior of the duplicate detection tool with the number of false negatives when tested on the large GeoNames dataset. Meanwhile, the polygon data generation tool proved to be a viable tool for creating accurate spatial representation from reference points of places lacking it. While this tool achieved good results in the majority of scenarios there still exist some limitations to it primarily unbalanced area distribution of different sized regions and the accurate recreation of irregularly shaped polygons.

This project's gazetteer can also improve greatly from future additions and improvements to already existing functionalities. For example, new data export formats can be added to the web application interface or a better filter and search system for interacting with the gazetteer's database. Likewise, the completeness of the gazetteer can be improved by importing further relevant data upon which its data integration and export features can be utilized.

Finally, improvements to the quality of the data integration methods used are also possible mostly by adopting methods from related works of the same type. Regarding the duplicate detection methods, while it provided decent results overall, the use of

more complex methods for string matching such as Santos et al. (2018) neural network based string matching mechanism could further improve the results of this integration process. Meanwhile, the polygon generation tool could also be improved by tackling the limitations described in its results section (e.g. handling of adjacent different sized regions). A possible solutions would be to make use of contributions from related works in the area such as Jones et al. (2008) web search based vague place delineation or one-class supervised learning through photo images of Cunha and Martins (2014) related work.

References

Acheson, E., De Sabbata, S., and Purves, R. (2017). A quantitative analysis of global gazetteers: Patterns of coverage for common feature types. *Computers Environment and Urban Systems*, 64:309–320.

Berman, L. (2016). *Placing Names: Enriching and Integrating Gazetteers.* Indiana University Press.

Berman, L., Mostern, R., and Southall, H. (2011). On historical gazetteers. *International Journal of Humanities and Arts Computing*, 5:127–145.

Cunha, E. and Martins, B. (2014). Using one-class classifiers and multiple kernel learning for defining imprecise geographic regions. *International Journal of Geographical Information Science*, 28(11):2220–2241.

Harpring, P. (2010). Development of the getty vocabularies: AAT, TGN, ULAN, and CONA. *Art Documentation: Journal of the Art Libraries Society of North America*, 29(1):67–72.

Hill, L., Frew, J., and Zheng, Q. (1999). Geographic names. *D-Lib Magazine*, 5(1).

Jones, C. B., Purves, R. S., Clough, P. D., and Joho, H. (2008). Modelling vague places with knowledge from the Web. *International Journal of Geographical Information Science*, 22(10):1045–1065.

Manning, P. (2015). World-historical gazetteer. *Journal of World-Historical Information*, 2.

Purves, R. S., Winter, S., and Kuhn, W. (2019). Places in information science. *Journal of the Association for Information Science and Technology*, 70(11):1173–1182.

Santos, R., Murrieta-Flores, P., Calado, P., and Martins, B. (2018). Toponym matching through deep neural networks. *International Journal of Geographical Information Science*, 32(2):324–348.

Southall, H., Aucott, P., Fleet, C., Pert, T., and Stoner, M. (2017). Gb1900: Engaging the public in very large scale gazetteer construction from the ordnance survey "county series" 1:10,560 mapping of great britain. *Journal of Map Geography Libraries*, 13(1):7–28.

Voronoi, G. (1908). Nouvelles applications des paramètres continus à la théorie des formes quadratiques. Deuxième mémoire. Recherches sur les parallélloèdres primitifs. *Journal für die reine und angewandte Mathematik (Crelles Journal)*, 1908(134):198–287.

Wick, M. and Vatant, B. (2012). The geonames geographical database. *Available from World Wide Web: http://geonames.org.* Accessed: December 2020.