Data Integration in Support of Gazetteer Development

Bruno Filipe Braz Magalhães

Thesis to obtain the Master of Science Degree in

Information Systems and Computer Engineering

Supervisors: Prof. Bruno Emanuel Da Graça Martins
Dr. Jacinto Paulo Simões Estima

Examination Committee

Chairperson: Prof. David Manuel Martins de Matos
Supervisor: Prof. Bruno Emanuel Da Graça Martins
Member of the Committee: Prof. Armanda Rodrigues

January 2021

Acknowledgments

I would like to thank my family for supporting me trough out the years and allowing me to reach this point. Without their incredible caring and encouragement I wouldn't have been able to complete this project.

Likewise, I would like to thank my close friends for heir friendship and help which allowed me to overcome some of the many hurdles and challenges this project presented.

Lastly but not least, I would also like to acknowledge the help and guidance of my supervisors Prof. Bruno Martins and Prof. Jacinto Estima. Their knowledge and insight on the topic of this research project provided an invaluable help during the development of this thesis.

To all the people mentioned above - Thank you.

Abstract

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 database management of a gazetteer and the challenges of data 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 for an historical gazetteer was developed utilizing contributions from other related projects such as the Alexandria Digital Library and the World Historical Gazetteer projects. The final gazetteer was then 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 distribution 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 a polygon data generation tool. 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 vague places.

Keywords

Historical Gazetteer, Places, Data Integration, Geometry Generation, String Matching, Visual Interface, Data Export

Resumo

Gazetteers fornecem informação de locais geográficos e as suas respetivas representações espaciais e relações. Ao longo do tempo, vários projetos procuraram desenvolver formatos e estruturas capazes de gerir este tipo de dados. A criação da base de dados de um gazetteer envolve também a implementação de métodos capazes de integrar dados de diferentes fontes externas, com modelos de dados distintos, e a delineação de regiões espaço-temporais para os dados importados. No entanto, atualmente não existe um projeto com uma solução universal para a gestão e integração de dados neste contexto de desenvolvimento de gazetteers. Esta dissertação tem como objetivo criar um gazetteer com uma base de dados funcional e ferramentas especificamente desenvolvidas para facilitar o processo de integração de dados. Assim, foi desenvolvido um gazetteer, utilizando contribuições de outros projetos relevantes, que será usado como base para a implementação de ferramentas relacionadas com integração e exportação de dados. O gazetteer foi disponibilizado a partir de uma aplicação web onde o utilizador pode interagir com a base de dados do gazetteer e exportar dados para um dos formatos disponíveis. Esta base de dados contém dados geográficos com foco em locais do periodo colonial Mexicano. A qualidade e vericidade destes dados foi avaliada de acordo com modelos para avaliação de gazetteers e a distribuição dos dados em questão. Esta avaliação confirmou a consistência e vericidade do gazetteer quando comparado com outras fontes relevantes. As principais ferramentas do processo de integração de dados consistem num detetor de duplicados e num gerador de poligonos sendo ambas avaliadas de acordo com cenários especificos para esse efeito. Os resultados demonstraram a capacidade de filtrar um elevado numero de duplicados e de gerar regiões espaciais realistas, respetivamente.

Palavras Chave

Gazetteer Histórico, Locais, Integração de Dados, Geração de Geometrias, Correspondência de nomes, Interface Visual, Exportação de Dados

Contents

List of Figures

List of Tables

Listings

1

Introduction

Contents

1.1 Context and Motivation

Geography plays a crucial role in many scientific fields and political decision making by providing information on places and their spatial location and relationships. This type of information can be stored in a gazetteer and afterwards represented in more well known geographic outputs such as maps or atlases. In particular, a gazetteer corresponds to a database storing information on places [3], each with a stable identifier and a number of descriptive properties (e.g., alternative names and the physical location).

A gazetteer database provides useful information for a wide variety of projects that deal with geographical information. This provides a way to locate places based on their name, answering ”where is” type of questions (i.e. where is London?). Additionally, gazetteers also store information on alternative names for places, including the historical evolution of names, further expanding the ability to locate places based on their designated name [4]. Another function of gazetteers is the translation between a place’s name and its respective spatial shape, also referred to as geospatial footprint, which also allows the identification of specific geographical patterns and features in a designated area [5].

In addition to storing information on individual places, gazetteers also store information on their relationships such as the spatial relation between places such as a place contained inside another (e.g. Berlin is contained in Germany) or changes across historical periods for borders of places like countries or other administrative regions. Furthermore, places can also be associated with each other by their type, through the use of hierarchies, identifying where each place belongs in the overall context of the gazetteer.

The development of large coverage gazetteers typically requires data integration and extraction from a variety of sources, a task that involves many data management challenges [6]. In particular, detection of possible duplicates when using multiple sources, vague location definitions, contradictory information on the same place and converting the input formats into the format used by the gazetteer.

Despite the prevalence of these issues across multiple gazetteer projects there is still a lack of a single gazetteer framework with a focus on handling this type of data integration challenges. Therefore, there was value in the creation of a framework which handles these issues in an efficient manner. Such framework should act as a data integration tool as well as a source for all kinds of place data for other gazetteer projects. The lack of this type of framework provided the main motivation for this research project and represents the main objective of this entire thesis project.

1.2 Thesis Proposal

The main objective at the start of this research project was the creation of an historical gazetteer that combines the positive contributions from previous gazetteer projects and related works. In particular, a historical gazetteer capable of managing gazetteer data, supporting data integration from multiple data sources, and estimate spatial and temporal data (i.e. regions and timespans). Furthermore, a web application was also developed in parallel to allow the exploration and representation of the gazetteer. To this end, the following objectives were established and fulfilled prior to the writing of this master thesis document:

1. Analysis of data models and methodologies proposed in previous projects related to gazetteer development, specifically the Alexandria Digital Library Gazetteer Content Standard [7] and the Linked Places format used in the World Historical Gazetteer [8], in order to identify their contributions.

2. Development of a database and web application for supporting the management of gazetteer data which made use of the formats and methods explored in the previous point.

3. Implementation and subsequent evaluation of data integration procedures for conflating gazetteer data gathered from multiple sources (e.g. duplicate detection methods and spatial data generation).

1.3 Document Organization

This thesis is is organized as follows: Chapter 1 provides a general introduction to the topic and the definition of the main objectives. Chapter 2 focuses on describing some necessary concepts and theories which were the basis for both this thesis as well as the relevant related works used as contributions to this one. Following this note, Chapter 3 presents all other gazetteer related works whose contributions were instrumental to the creation of the historical gazetteer framework reflected in this document. Chapter 4 describes the structure and functionalities of the developed historical gazetteer as well as the development process of the gazetteer itself in detail. With the base framework and its development specified, Chapter 5 then focus on presenting the final results and data of the resulting historical gazetteer. Of course, these results are also evaluated according to predetermined evaluation methods. Finally, Chapter 6 provides a final conclusion to the entire thesis with the mention to possible future work and improvements.

2

Fundamental Concepts

Contents

Building a gazetteer requires knowledge from multiple areas, ranging from how to represent the data to methods for creating regions. This section presents some essential concepts related to the development of the gazetteer database. First, we start by giving a small introduction to data management technology with a focus on techniques useful for modelling and representing gazetteer data (Section 2.1). Second, we provide a general description of various string matching techniques (Section 2.2) useful for detecting duplicate places in a gazetteer. Finally, we deal with methods for the delineation of regions based on multiple coordinates, useful for cases where a place's spatial region is vague or undefined (Section 2.3).

2.1 Data Management Technology and Linked Open Data

The origins of data management technology date back to the 1980s, with the transition from sequential data access to random store access and later on real time interactive usage. At its core, data management technology refers to all the techniques used to organize and store data in database systems. The main goal of this process is to allow the retrieval of information from database systems at any point in time.

Linked Data is nowadays considered to be one of the core pillars of information sharing on the Web. In other words, Linked Data provides a straightforward way for creating links between datasets that are designed to be easily understandable by machines. This domain of connections is often denominated as the Semantic Web, which itself is an extension of the World Wide Web (WWW) created in order to allow for a better exchange of metadata between software programs.

The creator of the WWW and a main advocate for the Semantic Web Sir Tim Berners-Lee defined in 2006 [1] the 4 main rules that form the basis behind Linked Data:

1. Uniform Resource Identifier (URI) should be used for naming
2. Use HTTP URI to allow the search of each name
3. Provide useful information using standards when a URI is looked up (RDF, XML, etc)
4. Include links to other URI to allow further discovery

The concept of an open WWW also created the necessity for data to be publicly available to anyone. This type of data is commonly named Open Data and refers to when data is made freely available to be used and redistributed by anyone without restrictions like copyright, access or patents. While still a relatively new concept, Open Data serves as a tool to bypass conventional ideas of owning and rights from being applied to data in the public domain.

[1] https://www.w3.org/DesignIssues/LinkedData.html

Linked Open Data combines both the linked machine readable data from Linked Data with the open source nature of Open Data. This creates a powerful tool for information gathering and storing that, due to the use of standard readable data formats, can be used across many sources with different formats. In the context of this project Linked Open Data properties make it a useful and simple data format to be used in a gazetteer database system. There are many types of Linked Open Data formats, based upon widely used languages for data representation such as JSON and XML, that can be used to represent geographical data.

A widely used Linked Open Data format is JSON-LD that makes use of the JSON making it an ideal option for programming and database usage. A variant of this linked data format specifically for geographic data storing is GeoJSON which also allows for the representation of geometric types.

Liewise, XML (Extensive Markup Language) is a spatial data representation format that combines both machine and human readability. The simple and general nature of XML allows it to be used across a multitude of languages, for example in the representation of data structures or the contents of Web Services. Many specifications based around the standard XML format exist, including formats specifically created to be used in scenarios involving geographic data. A relevant variant is GML (Geographic Markup Language) defined in 2000 by the Open Geospatial Consertium (OGC) [9] , whose main goal is to represent geographical data including both points and shapes. Another markup language used for geographical data representation is WKT (Well Known Text), also developed by the OGC, that focus on representing geometrical objects with their string notation. Some examples of WKT for representing geometric objects (geometrical shapes for regions are explored in more detail in Section 2.3) are present in the list bellow:

- **Point:** *POINT (2 5)*;
- **Line:** *LINESTRING (2 5, 10 4, 5 7)*;
- **Polygon:** *POLYGON (2 5, 10 4, 5 7, 2 5)*;

2.2 String Similarity Metrics

The topic of string similarity has been the subject of a large number of studies [10]. The main purpose behind this topic is to find effective ways to compare strings using the distance or similarity between them. The methods for calculating string similarities can be subdivided into 3 different main types: character based distance, vector based distance and hybrid approach.

Character based methods calculate the distance between 2 strings by taking into account the necessary operations to transform one string into the other. Such operations, much like the name of the method implies, are based on character editions such as deletions, substitutions and insertions. One of the most well known methods is the Levenshtein distance [11], originally envisioned by the Soviet mathematician with the same name. The Levenshtein distance is reached by calculating the minimum number of the aforementioned character operations in order to transform one string to the other. Over the years multiple variants of the original Levenshtein distance were proposed, such as the algorithm proposed by Damerau [12] that adds the transposition of neighboring characters as an operation. The Damerau-Levenshtein distance can be computed through a dynamic programming algorithm presented in Equation 2.1. The similarity metric (s) between sentences *a* and *b* is obtained according to: $s_{a,b} = 1 - d_{a,b}$. Where, $d_{a,b}$ is the distance between sentences *a* and *b*.

$$d_{a,b}(i,j) = \begin{cases} max(i,j) & \textbf{if } min(i,j) = 0 \\ min \begin{cases} d_{a,b}(i-1,j)+1 \\ d_{a,b}(i,j-1)+1 \\ d_{a,b}(i-1,j-1)+1_{a_i \neq b_j} \\ d_{a,b}(i-2,j-2)+1 \end{cases} & \begin{array}{l} \textbf{if } i,j > 1 \\ \textbf{and } a_i = b_{j-1} \\ \textbf{and } a_{i-1} = b_j \end{array} \\ min \begin{cases} d_{a,b}(i-1,j)+1 \\ d_{a,b}(i,j-1)+1 \\ d_{a,b}(i-1,j-1)+1_{a_i \neq b_j} \end{cases} & \textbf{otherwise} \end{cases} . \tag{2.1}$$

Another character based method is the Jaro similarity metric [13] which is primarily used for name matching since it takes into account the order of characters and the length of the strings being compared. On Equation 2.2 we define the Jaro similarity metric with *m* being the number of matching characters between the two strings *a* and *b*, and *t* being half the number of character transpositions.

$$s_{a,b} = \begin{cases} 0 & \textbf{if } a = 1 \\ \frac{1}{3}\left(\frac{m}{|a|} + \frac{m}{|b|} + \frac{m-t}{m}\right) & \textbf{otherwise} \end{cases} . \tag{2.2}$$

For the two strings to be considered a match, the following condition must be true:

$$s_{a,b} <= \frac{\text{max}(|a|,|b|)}{2} - 1. \tag{2.3}$$

However, the most used variant of the Jaro method is the Jaro-Wrinkle approach. Proposed by Wrinkle [14], the Jaro-Wrinkle algorithm is an evolution of the original Jaro metric by favouring strings that match from a certain initial set prefix. The Jaro-Wrinkle value (represented in Equation 2.4) corresponds to the sum between the Jaro value between the 2 strings and the following product. In this equation δ represents the size of the similar prefix at the start of both strings with a set maximum of 4 characters and *p* is a preset constant scaling factor that should be bellow 0.25.

$$s_{a,b} = s'_{a,b} + (\delta \times p \times (1 - s'_{a,b})). \tag{2.4}$$

Meanwhile, vector based approaches convert the strings in question into a corresponding vector representation and calculate the similarity value between those resulting vectors. The cosine similarity is a common method in this type of string matching that calculates the distance between 2 strings through the cosine of the vectors that represent them. More formally, the cosine similarity between vectors A and B can be defined as:

$$s_{a,b} = \cos(\theta) = \frac{A \cdot B}{||A|| \times ||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i} \times \sqrt{\sum_{i=1}^{n} B_i}}. \tag{2.5}$$

Another existing vector based string matching method is the Jaccard similarity coefficient [15]. This a simple method where the similarity between strings is the ratio between the intersection and the union of both strings. The Jaccard similarity is defined in the following equation where A and B are vectorial representations of 2 separate strings:

$$s_{a,b} = \frac{|A \cap B|}{|A \cup B|}. \tag{2.6}$$

Finally, hybrid approaches attempt to combine the advantages of both character based and vector based methods by striving to allow some difference between individual words and more flexibility in the word order. Hybrid methods are based around sub measures (represented by $s'_{a,b}$) between all the elements of both strings and calculating a final similarity value through these sub measures. An example of an hybrid method is the following equation proposed by Monge [16] that uses a sub measure based on the Jaro-Wrinkler algorithm:

$$s_{a,b} = \frac{1}{|a|} \sum_{i=1}^{|a|} \textbf{max}_{j=1}^{|b|} s'_{a_i,b_j}. \tag{2.7}$$

In the equation above, $s'_{a,b}$ is the Jaro-Winkler similarity value between an element a_i from string *a* and the element b_j from string *b*, with $|a|$ and $|b|$ being the length of both strings *a* and *b*, respectively.

Hybrid approaches can also be used in combination with the vector based procedures explored previously in this section such as the cosine similarity and the Jaccard similarity. Following up, in Equation 2.8 and Equation 2.9 show a variant of the previous hybrid algorithm from Morge and Elkan:

$$s_{a,b} = \frac{Z}{|a| + |b| - Z}, \tag{2.8}$$

with:

$$Z = \frac{\sum_{i=1}^{|a|} \textbf{max}_{j=1}^{|b|} (s'a_i, b_j) + \sum_{i=1}^{|b|} \textbf{max}_{j=1}^{|a|} (s'a_i, b_j)}{2}. \tag{2.9}$$

In those equations, $s'_{a,b}$ represents the Jaro-Winkler value of an element a_i from the string *a* and an element b_j from the string *b*. However, a new variable Z is present that corresponds to the similarity matching of elements of string a with string b and vice versa. The method is completed through Equation 2.8 which calculates a Jaccard similarity value from the Z obtained previously.

Some other approaches to string matching techniques do exist, such as phonetic based encoding methods that attempt to encode a string's pronunciation. However, such methods are not as reliable as the previously discussed domains of character based, vector based and hybrid approaches and as such will not be adressed in this project.

2.3 Region Delineation

In a geographical context a region is composed by a set of locations with at least one common attribute that connects them together. The creation of regions can vary from human attributes such as nationality and culture (i.e. countries or states) to those of natural origin (i.e., mountains or continents). Outside the geographical context, regions play an important role in facilitating governmental activities such as territorial and population management. The division of large areas into smaller regions allows for the implementation of measures specific to each region. Therefore, regions and subsequently their representation are an essential aspect of a gazetteer database, such as the one described in the context of this project.

A gazetteer database is composed of many geographical entities/locations with their own specific attributes and properties. Every named place has a corresponding physical location represented by one or more coordinates. This location can be represented by a single pair of coordinates or a complex region with multiple pairs of coordinates associated to it. Entries with only a single pair of coordinates can easily be represented by a simple point corresponding to that coordinate's location. The physical location of a place can also be represented by an array of coordinates that when connected can form more complex shapes such as lines and polygons. However, issues arise when the set of coordinates does not correspond to a predetermined geometrical shape (i.e. random cluster of points and/or lines).

Computational geometry deals with the creation and study of algorithms within a geometrical context and application. The recent developments in computation capacity and computer graphics has resulted in the rapid evolution of this particular field. The scope of computational geometry ranges from simple projections on a 2D space to the creation of 3D objects and environments or even applications in some other scientific fields such as robotics and circuit design. Most algorithms in computational geometry strive to be efficient and as such must take into account their complexity making them ideal for use in large datasets. For the purposes of this project, the remaining of this section will focus on describing some geometrical computation methods that will be used to address the previously established issue of region creation through a set of random points.

Given a random set of points, the resulting convex hull is the minimal convex shape that contains all those points. This shape can be reached by intersecting all possible convex combinations in the points

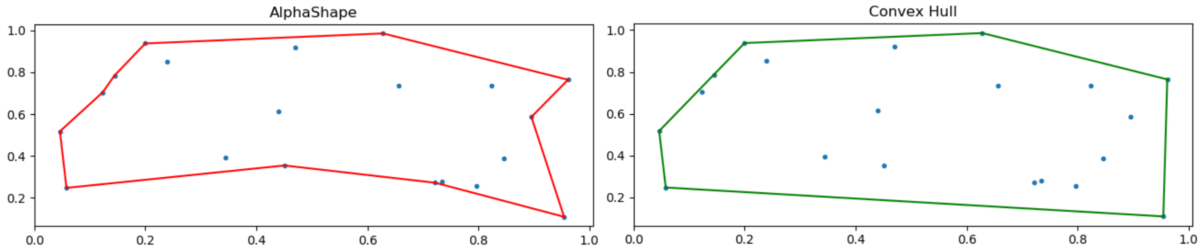


Figure 2.1: The resulting alpha shape and convex hull from the same set of points.

of the set. However, the resulting convex hull is still composed of a significant portion of empty space that corresponds to none of the original points. Alpha shapes [17], also referred to as concave hulls, are a generalization of the concept of convex hulls that look to diminish this empty space by more accurately representing the shape of the set of points. Given a random set of points, the resulting alpha shape will be a variant of the Delaunay Triangulation [18] of that set of points. The shape is determined by a preset constant α that follows the following definition:

- if $\alpha < 0,$ a closed disk of radius $\frac{1}{\alpha}$ is used
- if $\alpha = 0,$ a halfplane is used
- if $\alpha > 0,$ a complement of closed disk of radius $-\frac{1}{\alpha}$ is used

The algorithm for calculating the alpha shape consists on building a graph where every point in the set is a vertex. Afterwards, edges are built between every pair of points that are within the same disk radius previously established. Figure 2.1 showcases the differences between an alpha shape representation and a more simple convex hull representation, respectively.

Another possible method for region delineation is through the use of Voronoi diagrams [19] which are also based on the Delaunay Triangulation. Figure 2.2 presents the resulting diagram from a set of points on a 2D space. A Voronoi diagram (represented by a group of regions *R*) is created by attributing a region to all the points of the 2D space based on which point of *S* they are closest to. This is defined in Equation 2.10 where *p* is a single point in the 2D space where *S* is located and *d(x,y)* represents the distance between points *x* and *y*.

$$R_i = d(p, S_i) <= d(p, S_j), \textbf{where } i \neq j \tag{2.10}$$

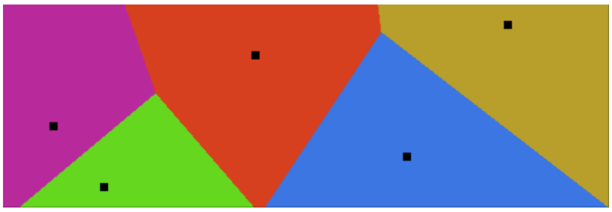


Figure 2.2: The resulting Voronoi Diagram from a random set of points.

In the context of this research project, a more complex method of region delineation that extend the base concept of Voronoi diagrams was developed to determine regions for places lacking one. Said method utilized the methodology used to create a Voronoi diagram with some additions to improve its overall accuracy and usability.

3

Related Work

Contents

This chapter starts by describing previously developed gazetteer projects and the methodology behind their development. The remaining sections explore specific works of data integration and data extraction in the context of creating digital gazetteers.

3.1 Gazetteer Development

Initial digital gazetteer development dates back to the 90's, with these digital variants being proposed as an alternative to traditional physical gazetteers. Over time the methods and capabilities of digital gazetteers evolved alongside technological advancements, most notably the rise of the internet as a source of information. This subsection present some gazetteer projects in detail which are considered relevant in the context of this thesis' historical gazetteer.

3.1.1 Alexandria Digital Library Gazetteer

The Alexandria Digital Library (ADL) [7] is a digital library manly focused on storing and organizing referenced geographical data that can be accessed through an HTML client. The origins of this project dates back to 1994, with its main goal being the creation of a digital alternative to the traditional extensive map library collection from the university of Santa Barbara in California. Some of the problems that affect traditional map collections range from simple access and storage issues to more complex ones like cataloging and finding the respective map for a specific location. The ADL gazetteer initially combined gazetteer information, available at the time of its inception, from federal agencies (the *United States National Imagery and Mapping Agency* and the *United States Geological Survey*) but has since evolved and is now available as a web service to users worldwide [1].

However, the transition from a physical to a digital storage requires a new type of structural organization capable of handling the gazetteer data. The ADL redefines the notion of a place in order to better suit the context of a digital gazetteer. Therefore, in the ADL gazetteer, a place is composed of the following 3 main elements:

- **Placename:** The term used to identify a place as well as any possible variant or alternative designations. Each place is only required to have one name, however a single place can have additional names as either alternative or historical designations. It is also possible to identify attributes related to each name such as the origin and pronunciation.
- **Placetype:** This relates to the classification of the nature of the place according to a list of predefined categories and types. It allows to more easily contextualize places and their respective

[1] http://www.alexandria.ucsb.edu/

relations and similarities among each of them. In the ADL the type of a place can also be derived from analysing the place's name (e.g. *Lake Michigan* will be classified as a *lake* type).

- **Footprint:** Corresponds to the physical location of a place which in turn is represented by a geometry (i.e. point, line, polygon, etc). The dimensions of each geometry are represented through the coordinates that correspond to the maximum scope of the geographical feature. The footprint of place can also be a collection of geometries if necessary.

Additionally, in order to handle historical data, a time period is attributed to all the elements described above or even the overall place itself. This temporal description corresponds to the time period of when a particular feature was valid. Furthermore, all the elements of a place also include metadata which stores links to external sources and contributors.

After an initial period of experimentation with this concept of place, the ADL gazetteer was remade following a formal structure referred to as the ADL Gazetteer Content Standard (GCS)[2]. The GCS structure's main goal was to provide a framework for recording named geographic places according to the aforementioned ADL place definition. Thus, the GCS supports both current and historical data for each place in the database. Overall the GCS is a large archive that associates named places with their locations, alternative names or even less concrete relational queries (e.g. What are the hospitals in Lisbon?).

The GCS makes use of an XML schema as a basis for the creation of a relational database. The XML schema is comprised of all the necessary attributes to describe a place.

With the GCS allowing for the classification of places into specific types it was necessary to develop a system for describing places and features. The solution was the creation of the ADL Feature Type Thesaurus (FTT), a hierarchy of multiple terms used to label each place in the gazetteer and their features. The FTT was built following the standard thesaurus conventions and its hierarchy uses terms from multiple sources such as the National Imagery and Mapping Agency or the Canadian Permanent Committee on Geographic Names.

A simplified version of the GCS relational database schema is represented in Figure 3.1. As one of the earlier digital gazetteer projects, this database model provides a solid base template for the construction of a gazetteer database. Therefore, the GCS model is something to take into consideration for the purposes of this project since it proves to be a very stable and effective foundation to start building a digital gazetteer database. It still suffers from some limitations as one of the early developed projects (i.e lack of effective integration tools between gazetteers and relations between places) which would be tackled in future developments.

[2] http://legacy.alexandria.ucsb.edu/gazetteer/ContentStandard/version3.2/GCS3.2-guide.htm

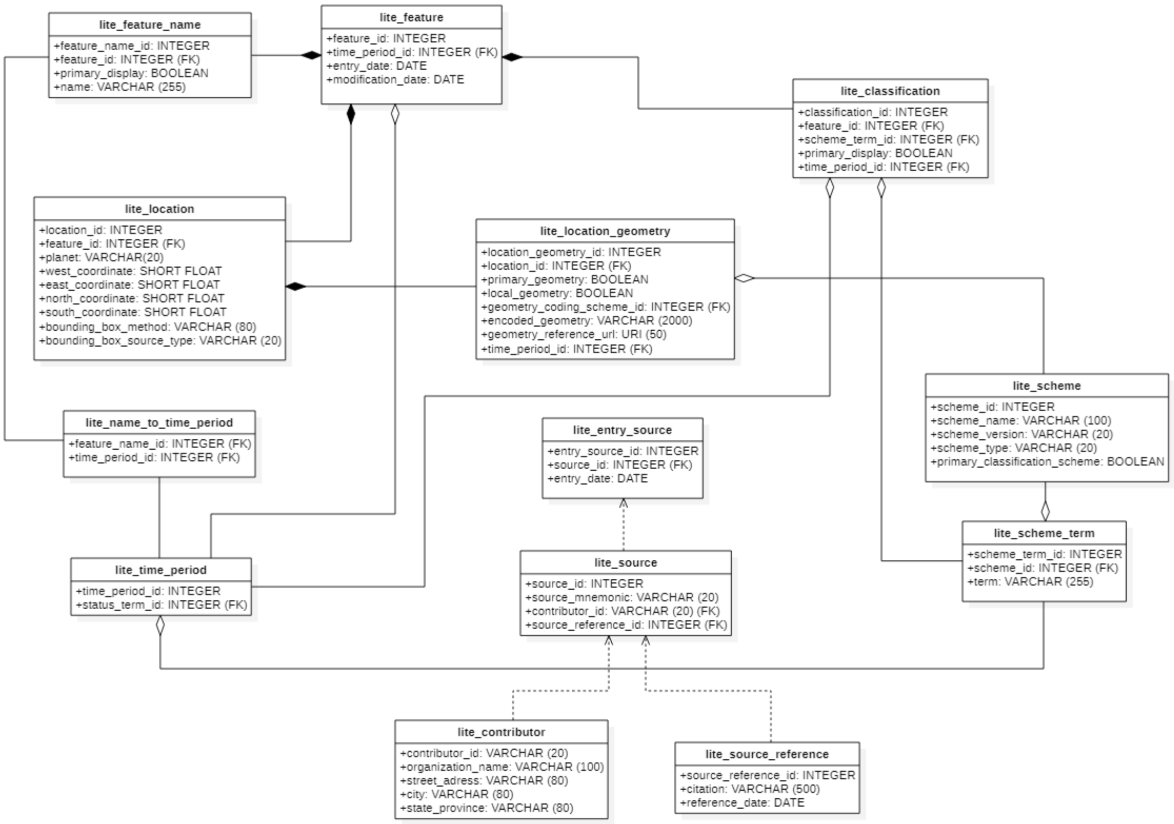

lite_feature_name
+feature_name_id: INTEGER
+feature_id: INTEGER (FK)
+primary_display: BOOLEAN
+name: VARCHAR (255)
lite_feature
+feature_id: INTEGER
+time_period_id: INTEGER (FK)
+entry_date: DATE
+modification_date: DATE
lite_classification
+classification_id: INTEGER
+feature_id: INTEGER (FK)
+scheme_term_id: INTEGER (FK)
+primary_display: BOOLEAN
+time_period_id: INTEGER (FK)
lite_location
+location_id: INTEGER
+feature_id: INTEGER (FK)
+planet: VARCHAR(20)
+west_coordinate: SHORT FLOAT
+east_coordinate: SHORT FLOAT
+north_coordinate: SHORT FLOAT
+south_coordinate: SHORT FLOAT
+bounding_box_method: VARCHAR (80)
+bounding_box_source_type: VARCHAR (20)
lite_location_geometry
+location_geometry_id: INTEGER
+location_id: INTEGER (FK)
+primary_geometry: BOOLEAN
+local_geometry: BOOLEAN
+geometry_coding_scheme_id: INTEGER (FK)
+encoded_geometry: VARCHAR (2000)
+geometry_reference_url: URI (50)
+time_period_id: INTEGER (FK)
lite_scheme
+scheme_id: INTEGER
+scheme_name: VARCHAR (100)
+scheme_version: VARCHAR (20)
+scheme_type: VARCHAR (20)
+primary_classification_scheme: BOOLEAN
lite_name_to_time_period
+feature_name_id: INTEGER (FK)
+time_period_id: INTEGER (FK)
lite_entry_source
+entry_source_id: INTEGER
+source_id: INTEGER (FK)
+entry_date: DATE
lite_scheme_term
+scheme_term_id: INTEGER
+scheme_id: INTEGER (FK)
+term: VARCHAR (255)
lite_time_period
+time_period_id: INTEGER
+status_term_id: INTEGER (FK)
lite_source
+source_id: INTEGER
+source_mnemonic: VARCHAR (20)
+contributor_id: VARCHAR (20) (FK)
+source_reference_id: INTEGER (FK)
lite_contributor
+contributor_id: VARCHAR (20)
+organization_name: VARCHAR (100)
+street_adress: VARCHAR (80)
+city: VARCHAR (80)
+state_province: VARCHAR (80)
lite_source_reference
+source_reference_id: INTEGER
+citation: VARCHAR (500)
+reference_date: DATE

Figure 3.1: Simplified version of the Alexandria Digital Library relational database diagram.

3.1.2 Getty Thesaurus of Geographical Names

Getty Vocabularies [20] are a group of vocabularies that compile different terminologies related to art, architecture and cultural domains. Getty vocabularies are all interconnected with each other and follow international standards for data presentation when providing information about their respective domains. Getty vocabularies also support multilingual names in both current and historical terms. Among these vocabularies is the Getty Thesaurus of Geographical Names (TGN) whose main purpose is to provide information about current and historical places that are relevant in the context of art, architecture and other cultural domains present in the other vocabularies.

The Getty TGN focus on the hierarchy and both internal and external relations of all the terms in its vocabulary, therefore it should not be confused with a conventional geographical tool (i.e. maps) since it lacks the geographical representation of the places listed. Some records do include coordinates to their location however these are just approximations and should be used only as broad references. Being a thesaurus, TGN attempts to compile all the synonym names that refer to the same entity into a single record. The scope of the TGN includes places from all continents of the world, even some extraterrestrial locations, making use of the multilingual support of the Getty vocabularies in order to accurately gather as many names associated to a place as possible. Lastly, the types of place in the TGN can be both of political nature, such as countries or regions, or physical nature referring to geographical entities like mountains or rivers.

To effectively handle data on such a large scale, the TGN makes use of a data model that facilitates the use of multiple names and allows to easily establish hierarchical relation between different terms. All records have a unique and persistent id (represented by the field *SubjectID*) which guarantees that there aren't 2 records referring to the same entity. Each record is also attributed a type that reflects the nature of the place associated with it and can be one of the following:

- **Administrative entity:** Representing nations, states and any entity defined by human boundaries;
- **Physical feature:** Representing geographical elements such as mountains, rivers or any other types of natural features;
- **Both:** Representing entities that fall under both previous types (although this is a rare case);
- **Guide Term:** Used to facilitate hierarchical organization, mostly associated with historical places;
- **Facet:** Used primarily in high level hierarchical entities, including records like *World* and *Extraterrestrial places*;

Names in the TGN are used to identify a place which also includes alternative designations. Each name has a language of origin and can be either historical or current depending on the status of its usage. Just like records, every name as a unique id associated to it (i.e., *TermID*). A single place can have multiple names associated to it, however all names (historical and current) in a record must be true synonyms. This means a record can only have names that specifically designate the exact place of the record. Lastly, every place with multiple names has one of his names flagged as the preferred term which means that such name is the most commonly used designation.

The hierarchical feature traces back the record all the way to the Facet. Both administrative and physical records have hierarchical relations that help contextualize the place in the overall TGN vocabulary. Every new record in the TGN should have a close parent record associated to it (e.g., the closest parent to the municipality of Lisbon would be the larger district of Lisbon). Physical places may be associated with multiple parents and as such they can have additional hierarchical sections (i.e., the alps run across multiple countries with each of them being a possible parent).

All the places in the TGN have a place type to better categorize and understand their existence. Place types are subdivisions of the more abstract record types. For example, administrative type records can have *inhabited place* or *city* as their place type while physical type records can have *river* or *mountains*.

Just like names, records can have multiple place types (with a minimum of 1) being one of them flagged as the preferred type.

Finally, no place in the TGN exists in a vacuum which means that places can have relationships with other places. This is represented in the *Related geographic places* feature which lists all other records associated with the place in question.

3.1.3 World Historical Gazetteer

Although still relatively recent and in the middle of its development, the World Historical Gazetteer [8] (WHG) project makes use of some interesting methods in the context of this research project. During a 3 year period spanning between 2017 and 2020, the WHG project was focused on the gathering of data, mostly related to historical places, and their respective names and relations, for the purpose of building a gazetteer database. Furthermore, alongside the creation of the gazetteer database, the WHG project looks to develop a graphical interface and various programming tools to facilitate both the use of the WHG and future contributions to help expanding its scope even after the 3 year development period. To achieve this goal the WHG project uses Linked Open Data (described in Section 2.1) as a basis for its data format.

The base architecture of the WHG takes heavy inspiration from the Peripleo project [21] whose main goal was to create a historical gazetteer focused on ancient Mediterranean locations. This architecture can be divided into 2 main elements, the first dealing with the collection and integration of data from multiple sources into the gazetteer database and the second element allowing the storing of multiple annotations that linking places in the gazetteer to external items such as places of interest or events. Despite the similarities in its data format and architecture, the WHG and Peripleo differ heavily in scope when it comes to both time and geographical coverage. The WHG is much larger in scope since it includes places from the entire globe and its timeline is not bounded to a specific historical period. Therefore, some additional information on places (i.e. historical routes, environmental information and regional relations) is included in order to support all the additional relations across time and space that result from the larger scope of the WHG.

The structure of the WHG is composed of a main data storage named the Spine where more then 20000 places are stored. The WHG is also composed of 2 indexes with the first being a union index from contributors and sources of gazetteer data, while second index stores annotations of items related to places stored in the Spine. All this data is stored in a relational PostgreSQL database for easy access and manipulation. Through these indexes it is possible to establish links with identifiers from other gazetteers, such as the previously explored Getty TGN. These connections allow easy access to alternative names and even opens the potential for data mining across multiple place data sources.

Regarding the data format, the WHG makes use of the Linked Places Format (LPF). This format was adopted as a replacement to the originally used Pelagios Gazetteer Interconnection Format [22] since it provides a number of significant advantages when handling historical data. The main goals of using the LPF include: search of terms across different gazetteers, information gathering to identify or differentiate places and linking data to the correct gazetteer through annotations with Uniform Resource Identifiers (URI). Thus, LPF offers a data format that facilitates the linking of information between different gazetteers and sources, an important feature in an environment where many gazetteers make use of distinct data formats. The LPF is considered a Linked Open Data format and as such it uses the JSON-LD as its basis, more specifically the subset implementation GeoJSON-T. The use of this universal format allows the LPF to be used by a vast array of web applications while also providing temporal description of names and places. The following Listing 3.1 is a generic record following the LPF structure:

Listing 3.1: Generic LPF Example

```
FeatureCollection {
     @context: "http://geojson.org/geojson-ld/geojson-context.jsonld",
     @context: "http://whgazetteer.org/models/whgazetteer.jsonld",
    features: [{
        type: Feature,
        geometry: {
            type: GeometryCollection,
            geometries: [
                {type: [Point|Line|Polygon],
                coordinates: [[x, y], ...],
                loc_attestations:
                 [{source, contributors: [], when: {}, certainty}],}
            ],
        },
        when: {#computed from the "when"s of name and loc attestations},
        properties: {
            name_attestations: [{name, language, when: {}, isPreferred,
                        source, certainty}],
            relations: [{relation, placeUri, when: {}, source,
            certainty}],
            snippets: [ {uri, language, description,
                        source, certainty}],
            links: [ {type: [exactMatch|closeMatch|seeAlso|seeFurther], uri,
                        [who, when, note]}],
            modifications: [{timestamp, who, note}], note: ""
```

```
                }
        }]
}
```

For the record identification the LPF uses a structure similar to the GCS and Getty's TGN formats (e.g., multiple designations, hierarchical relations and unique identifiers). It also provides temporal information through the *when* elements which associate a place feature to the respective valid time period. However, the LPF sets itself apart from this previous formats through the addition of links to other gazetteer entries and external items (seen in the *relations* and *links* elements of the above example). This record of external links connects the gazetteer with other sources of place data and in turn facilitates the integration process with other gazetteers if necessary.

All the data present in the WHG is made available through an Application Programming Interface (API) that returns filtered records in multiple formats such as GEOJSON-LDT, CSV and RDF. Furthermore, the development of a web application is also under way which will work as an interface for exploring the gazetteer's data.

3.1.4 Other Large Gazetteer Projects

Over the years many projects used some of the previous discussed gazetteers, such as the ADL model or the TGN multiple name sources, as a basis to build their own gazetteer database system. Meanwhile, the rise of the internet in the last decade effectively changed the way gazetteers are built and explored. Some of these projects provide unique pieces which are relevant to highlight, such as the GB1900 and the 'Who's on first?' projects.

The GB1900 [23] project main goal was to build a highly detailed historical gazetteer of Great Britain and make it available through the use of Linked Open Data. The level of detail present in the GB1900 project allows to include every possible similar small scale place (e.g. village) in Great Britain into its database, creating the largest and truly comprehensive historical gazetteer up to date. However, such level of detail requires overwhelming amounts of data to be collected, organized and inserted into the gazetteer's database.

The solution to this challenge, and also the main point of interest of the GB1900 project, was the use of user engagement to expand its place data. New entries into the gazetteer are crowd sourced from the user base, gathered by exposing the project to the public eye with external publicity. The growing number of volunteers directly led to a large number of new places being added to GB1900. In a short amount of time most of the places had already been registered, proving crowd sourcing to be a viable approach for creating highly detailed gazetteer databases.

Meanwhile, the Who's on First?[3] project is one of the largest compilations of geographical open data

[3] http://www.whosonfirst.org/

as well as licensed official data. However, in the context of this project the main point to highlight is the ”Who’s on First” visualization tool Spelunker. This tool provides a simple interface that allows the user to explore the places and data of the gazetteer. This includes all the information belonging to a specific place such as the possible denominations, the geometry with the corresponding position in the world map, the type and other related places.

One of the goals of this project was also the development of a web interface that facilitates the exploration of the data in the gazetteer. The Spelunker is a perfect example of a tool that achieves this goal, in turn providing a great basis for this project’s own web application interface.

3.2 Data Integration and Gazetteer Conflation

The previous section explored a few digital gazetteer projects, with various ideas to help building gazetteer systems. However, developing a new gazetteer can also involve gathering information from other already existing gazetteers. The lack of a universal model for gazetteer creation means that traditional data integration techniques are necessary in order to correctly combine data from different gazetteer sources. This integration process comes across a few challenges most notably on how to deal with conflating data between different gazetteers, which in practice means the detection of similar or overall redundant entries across all sources.

This type of integration challenge is unique since it includes comparing geographical regionsas well as place denominations and types. Among the proposed solutions to this situation is the work of Hastings [24] which tries to establish a general approach to record matching in a gazetteer. This approach decomposes a place in a gazetteer database into 3 main core elements: the footprint (geographical representation), the placetype and the placename (the core elements of the ADL explored in Section 3.1.1). The similarity between 2 places is the equivalent to the similarity between their core elements. These 3 similarity metrics vary between the value 0 and 1, therefore it is considered to exist conflation between 2 places when all these 3 similarities are close to 1. There are a few drawbacks to Hasting’s approach particularly the fact that the methods used for comparison are more general and therefore the results may not be extremely accurate in some integration scenarios. Other research works propose more accurate and exhaustive methods for comparing the core elements of a place. The remainder of this section will present comparison methods for each core element of a place.

3.2.1 Footprint

Frontiera [25] provides an insight into many possible methods used for footprint comparison. Footprints are first separated into the possible types of representation: single point, minimum bounding box, convex hull and alpha shape. The main goal of every spatial comparison process is to calculate the spatial similarity between each region, followed up by a logistic regression to calculate its spatial ranking (probability of relevance).

The first step in assessing spatial similarity is to gather all the footprints of the places relevant to the query in question. In the case of geographical representation using only points, an inverse distance weighted (IDW) function is enough to get a fairly accurate spatial similarity metric. However, when it comes to more complex representations such as minimum bounding boxes (MBBs) or convex hulls, more complex methods are necessary. The majority of these can be grouped into two major approaches: area of overlap or Hausdorff distance. Area of overlap is similar to the method used by Hasting and simply calculates how much the 2 regions intersect each other. Meanwhile, Hausdorff distance between 2 regions A and B refers to the maximum of the following 2 distances: (1) the maximum distance that any point in A is to the nearest point belonging to B or (2) the opposite maximum distance of any point of B is to the nearest point belonging to A. When it comes to results, a value of 1 for the area of overlap means a perfect similarity, while the opposite applies to the Hausdorff distance where a value of 0 means the 2 regions are identical.

Realistically, perfect similarity scores are a minority in the overall number of comparisons which means that the majority of comparisons end up with similarity values that are hard to interpret. Frontiera proposes a probabilistic model that makes use of the previously obtained similarity scores which represent the probability of a region being relevant to a query or similar to another region. The probabilities are modelled using a logistic regression represented in equation 3.2, estimating the relevance log-odds *R*. The pair of regions is described by a set of coefficients β alongside a set of *N* feature variables *X*,

$$\log O(P|R) = \beta_0 + \sum_{i=1}^{N} \beta_i X_i. \tag{3.1}$$

Finally, the resulting log-odds are transformed into probabilities of relevance, equivalent to the spatial ranking, through the following equation:

$$P(R|X) = \frac{1}{1 + e^{\log O(P|R)}}. \tag{3.2}$$

3.2.2 Placetype

Contrary to placenames and footprints, placetypes can have their base structure vary a lot between different gazetteer databases since each one can have their own placetype hierarchy. This variance makes it difficult to effectively compare places through their type across gazetteers with different hierarchical organizations. As such placetypes are the hardest core element of place to assert similarity in the context of integrating different gazetteer databases. Therefore, for the sake of simplicity the focus will be on comparison methods for placetypes within the same hierarchy.

The approach proposed by Hastings [24] in comparing placetypes within the same hierarchy is based around their distance in the hierarchical structure. For example, using the previously discussed Feature Type Thesaurus of the ADL as a hierarchy, the geographical types *cadastral areas* and *military areas* are both direct subtypes of *admnistrative areas*. The distance value is obtained through the sum of the necessary upwards steps (up-steps) in the hierarchy until the narrowest term common to both placetypes, which in the previous example means one up-step for each. This means that there is a distance value of two up-steps between the *cadastral areas* and *military areas* types. This distance is not linear since each up-step should be considered a massive variance between the types being compared, specially the more upwards their common element is in the hierarchy. Therefore, each up-step is considered an exponetial step in distance which in turn leads to the following similarity equation:

$$S = 2^{-(Upstep(T1,T2)+Upstep(T2,T1))}. \tag{3.3}$$

Given the T1 and T2 placetypes of the same type hierarchy, the similarity of these 2 terms is the result of the inverse up-step exponential values. As such in the case of complete mismatch the up-step will be infinite leading to a similarity of 0, with the opposite being true in the case of identical placetypes having each an up-step of 0 meaning a similarity of 1. To possibly apply this approach to placetypes in different hierarchies it is necessary to find a correlation between their respective type classifications.

3.2.3 Placename

Finally, placenames are represented as strings meaning that traditional string matching techniques can be used to compare and assert the level of similarity between placenames. However, in practice the distance or similarity resulting of applying string matching techniques to placenames may not correctly correspond to what they actually refer to in reality. To better illustrate this issue let's use the example of a city like New York. This entity has quite a few current denominations such as: New York, NY or NYC. If nicknames (e.g. The Big Apple) or historical names (e.g. New Amsterdam) are included as well the use of a string matching technique would not result in a high score of similarity, despite the fact that all of these effectively refer to the same place.

To try to circumvent this problem, Hasting's work utilizes a cross-matching algorithm based around stemming (the reduction of words to their linguistic root meaning) when dealing with gazetteer conflation. Instead of actually comparing both strings as a whole, this algorithm focus on matching each token (group of sequential characters separated by a white space or punctuation) between both strings disregarding their position on the actual string. For better accuracy stop words are also ignored in the entire algorithm. Each matching token is given a value of either 0, $\frac{1}{4}$, $\frac{1}{2}$ or 1 depending on the level of matching. Thus, the following similarity (*S*) equation for two strings *S1* and *S2* is achieved where *Match* represents the matching value and *Tokens* selects only the relevant tokens from the string to be compared:

$$S = \frac{Match(Tokens(S1), S2) + Match(Tokens(S2), S1)}{Tokens(S1) + Tokens(S2)}. \tag{3.4}$$

More recent research projects have tried to find more effective methods for string matching where placenames are also included into. A relevant work in the context of gazetteer integration is that of Santos et. al [1] which uses a supervised learning approach, more specifically neural networks, for the process of string matching. The main goal of this work was to detect duplicate entries including the case when a historical or alternative designation to refer to the same place. This process involves the use of classification based on parameters learned from data, instead of simple string matching algorithms.

A neural network is a composition of nested mathematical operations performed on the nodes of each layer. Individual nodes can also be referred to as perceptrons [26] with each perceptron receiving multiple weighted inputs and computing a single output through an output activation function. Equation 3.5 mathematically represents this calculation where y is the output prediction, **x** the vector of inputs, **w** the vector of weights, b the bias term and φ the activation function.

$$y = \varphi\left(\sum_{i=1}^{n} w_j x_i + b\right) = \varphi\left(\mathbf{w}^T\mathbf{x} + b\right). \tag{3.5}$$

A multi layer network is composed of an input set of nodes, a number of interior hidden layers and the output layer. In this type of network, also referred to as feed forward network, the input values propagate through the network layer by layer until it reaches the output layer where a single classification value is obtained. The training of such networks involves adapting the weight **w** values through a set of training inputs **x** taking into account the accuracy of the corresponding outputs.

Santos et. al approach, however, utilizes Recurrent Neural Networks (RNN), an extension of feed forward networks, which allow inputs of varying sizes. This is achieved through the use of a multi-step recurrent hidden state where each activation step is dependent on that of the previous step. There are two sources of input in a RNN: the present and the recent past. These are combined in a feedback loop of inputs and outputs which preserves the sequential information inside the hidden state. Given an input vector **X**, the RNN updates its recurrent hidden state by sequentially computing the following Equation 3.6 on the input set. Where **x** is the input at time **t**, **W** the weight matrix and **U** the transition

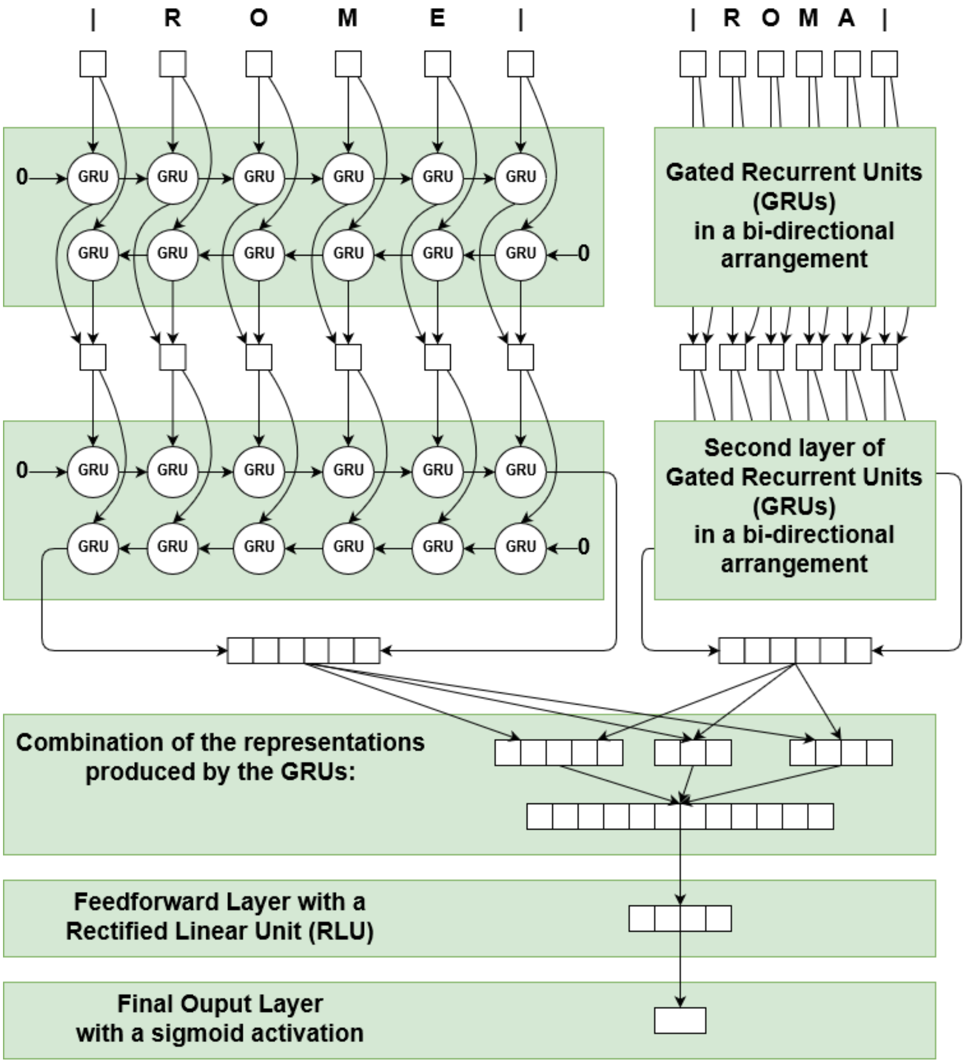

| R O M E |
| R O M A |
0
GRU
0
Gated Recurrent Units (GRUs) in a bi-directional arrangement
Second layer of Gated Recurrent Units (GRUs) in a bi-directional arrangement
Combination of the representations produced by the GRUs:
Feedforward Layer with a Rectified Linear Unit (RLU)
Final Ouput Layer with a sigmoid activation

Figure 3.2: The neural network architecture for string name matching. Modified from Santos et. al [1].

matrix from the previous state at time **t-1**.

$$\mathbf{h}_t = \varphi\left(\mathbf{W}\mathbf{x_t} + \mathbf{U}\mathbf{h_{t\text{-}1}}\right). \tag{3.6}$$

Therefore, the process for string matching proposed by Santos et. al, represented in Figure 3.2, can be summarized as follows:

1. Both strings are converted into a normalized unicode representation (i.e. UTF-8) and a special symbol is added to mark their beginning and ending. Afterwards, the strings in the unicode format are converted once again to a one hot binary vector representation.

2. Next step is to provide the binary vectors as input to the RNN, more specifically a subtype named Gated Recurrent Units (GRU) that better handles long input sequences. Being a RNN, the com-

puting process involves a multi layer bidirectional (left to right and right to left) computing of the input string vectors by the GRUs. The output will be a concatenated version of both directions real vector outputs, usually referred to as an embedding.

3. The two resulting outputs can now finally be compared and combined into a single representation through a set number of operations.

4. Finally, this combined representation is used as an input to a set of standard feed-forward neural network layers. The result will be a final output with the classification of the initial two strings as either matching or non-matching.

Note that, as a supervised learning classification method, the accuracy of string matching may vary with different parameters including the overall quality of the training data.

3.3 Data Extraction for Generating Gazetteers

The creation of a new gazetteer requires the gathering of a large set of data for all the places under the designated geographical scope. The aforementioned integration methods to combine data from other gazetteer projects may not provide all the necessary geographical data to fulfil the target scope of a new gazetteer. Therefore, additional proper methods for data extraction need to be taken into account, particularly those concerning data extraction from Internet sources. As the name implies, data extraction can be summarized as the process of data gathering from multiple sources and the subsequent operations to integrate that data into a specific database, in this case a gazetteer database.

This section will explore previous research on data extraction that focus on gathering and processing data for some of the core elements of a gazetteer database: general place features, vague region delineation and determining temporal designations.

3.3.1 Building Gazetteers from the Web

The creation of a complete and detailed gazetteer database requires the gathering and processing of a significant amount of data. Initial gazetteer projects relied on manually inserting geographical data from physical sources such as collection of maps from libraries or government organizations. Over time the growth and availability of the internet changed this reality. Data mining from web sources has become the main method of acquiring data that can reliably be used to create the places that populate a gazetteer database.

Goldberg [27] proposed one of the first solutions for effective web data extraction to be used in the creation of a gazetteer. The main goal of Goldberg's approach was to automatize the extraction of web data and afterwards generate places with detailed features to insert into the gazetteer database. The initial data extraction was achieved through the use of semi-supervised learning tools, more specifically using agents that extract data based on previous training examples. Agents can be used to extract structured data (i.e., JSON, XML) or extract features from the information present in the text of web pages. The accuracy and results of the agent output vary according to the quality of the sources used, with the ease of access and amount of data available being the major factors defining that quality. The algorithm used by the data extracting agent can be summarized in 3 steps, as follows:

1. The first step is to generate a set of generic geographical features that will be the basis for describing a place in the new gazetteer. Initially, a candidate set of features is generated based on urban addresses of the target geographical area. The existence of each generated address is then tested by an intelligent agent whose output is either a real address ID or an indication that the address is not valid. The result is a refined set of addresses that accurately represent the locations present in the geographical area.

2. The next step is to associate each address in the refined set to a name and a type that accurately describes it (this is done automatically if the information is available in the actual address). These associations are extracted from web sources (*Superpages* and *Switchboard*) by using the previously mentioned data extraction agents. The resulting names and types data are then normalized in order to be consistent with the ones generated by the gazetteer, alongside the removal of possible duplicate entries.

3. Finally, a spatial footprint is attributed to each of the places previously inserted into the gazetteer. This is achieved through the use of a geocoder specifically developed to determine the geographical location of addresses.

The rise of the social networks in the last decade brought with it a new way of gathering data for gazetteer development. However, it also created a whole new challenge to overcome on how to effectively extract and operate on these new large amounts of data from a computational level. These large amounts of data are commonly referred to as Big Data. In recent years, the change of the internet landscape prompted a necessity for an effective way to handle Big Data in the context of gazetteer development.

Gao [2] proposed a novel cloud-based solution to tackle the Big Data handling issue. The idea behind it is related to the creation of a scalable and distributed platform that can effectively generate gazetteer data from geographical Big Data Web sources [28]. This platform is named the Hadoop-based platform since the processing of Big Data is manly done through the cloud-computing platform *Apache Hadoop*. The Hadoop-based platform fulfils the requirements of scalability by distributing the computational work across a cluster of many small servers denominated the Hadoop-cluster. The Hadoop-cluster does

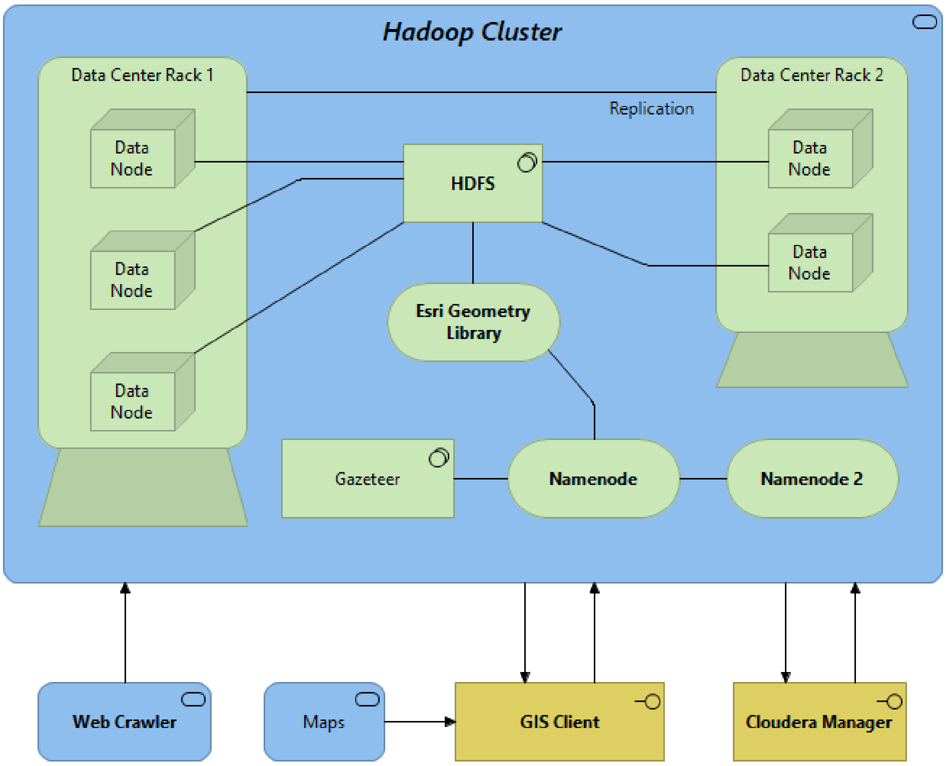


Figure 3.3: Hadoop system architecture [2].

not provide any type of visualization tools for statistics or mapping of the geographical data. Gao's alternative was the integration of both the ESRI Geometry API [4] working alongside a visualization app for data representation. The architecture of this platform is represented in Figure 3.3.

The first element that comes into play is the Web Crawler, a python search engine that extracts web place data and stores it on the main server. The web crawler can extract data from semi-structured data sources (social media) or unstructured text descriptions of places usually found in Web pages or documents. However, the primary element of the architecture is the Hadoop Cluster which itself is composed of two main components: the Hadoop distributed file system (HDFS) and the MapReduce programming tool. The large amount of processed data (structured or unstructured) is stored in the HDFS. The HDFS is composed by 3 types of nodes: *Name Nodes*, *Secondary name nodes* and *data nodes*. The name nodes deal with matters of metadata while the actual place data is stored in the data nodes in formats like XML or spreadsheets.

The MapReduce programming tool, which is a parallel process executed in the distributed servers of the Hadoop-cluster, allows the processing of the data extracted by the web crawler. The MapReduce decomposes a process into *map* and *reduce* sub-processes. The *map* sub-process begins with the division of the input by the *name node* server into smaller subsets. A key and value pair is given to each

[4] https://github.com/Esri/geometry-api-java

subset for identification and then the subsets are distributed across multiple *data nodes* for processing. The *reduce* procedure will then merge the output according to the previously given keys and send it back to the *name node*.

As stated before the analysis of data to identify place descriptions and features requires significant computational power. By diving the input into multiple smaller parts which are then concurrently processed, the MapReduce algorithm greatly mitigates the necessary computing time and effort.

It is also possible to perform spatial analysis in the Hadoop cluster through the Esri geometry library. Data is stored in a GEOJSON format on the HDFS in order to allow the handling of geometric features and operations. The Esri framework contains a geometric API from which spatial operations can be performed upon the geo data stored in the HDFS. These operations can be of two types: relationship analysis (e.g. equals, disjoints) or spatial operations (e.g. intersect, union).

Spatial analysis of extracted places also uses the MapReduce programming tool to allow the parallel spatial join operation of multiple places in the gazetteer. The main goal of this algorithm is to assign all places to the corresponding administrative area by analysing their spatial distribution. This algorithm makes use of two specific MapReduce functions:

- The *Mapper* which works similarly to the previous *map* function by assigning a key/value pair to each geographical feature, followed by checking if each feature contains the join target feature
- The *Reducer* that aggregates the results from the *Mapper* through a mapping operation

The Hadoop cluster activity is managed through the Cloudera Manager Web User Interface. The Cloudera Manager allows the deployment and management of operations of the entire Hadoop infrastructure. It also provides a real time information on the inner workings of the Hadoop cluster's nodes and services, also enabling configuration changes across the entire platform.

The final element of the architecture is a Geographic Information System (GIS) client that allows for the visualization of the results produced by the MapReduce programming section.

3.3.2 Delineating Vague Regions

So far places have been established as geographical entities whose main characteristics include a name, a type and a location represented by one or more coordinates. However, some places may not fall under this constraint in particular when it comes to having a well defined geographical area associated to it. Places with vaguely defined regions are usually not included in official gazetteer studies but they are commonly used to refer to certain geographical areas. Common referred places like *North of Portugal*, *the Caucasus* or *the American East Coast* are examples of places that lack an easily definable area.

It is important to take vague places into consideration when constructing a gazetteer, in order to identify their actual geographical region and their spatial relationship with other existing places. The main question then is how to actually define the region of a vague place, both the geographical footprint as well as how other places relate to it.

Jones [29] proposed a way to define the boundary of a vague region using the location of other places related to the vague place being defined. In other words, the main idea was to model a region based around which places are assumed to be a part of it or other places closely related to it. Web pages provide a vast and easy to use source for this purpose. Therefore, Jones method for vague region delineation can be summarized into four main parts: web search, extraction, coordinate assignment and boundary definition.

The first step involves the submission of search queries to the Google search engine containing the designation of the vague region to get information that can help defining it. Sometimes simply using the region's name may not prove sufficient so this method adds more details to the query like concepts typically related to that region or word patterns to better identify places within that region. Place names or other geographical features are then extracted from the top results. Named Entity Recognition (NER) methods were used to identify place names, in particular Jones work uses A Nearly-New Information Extraction (ANNIE) system, belonging to the General Architecture for Text Engineering. This system identifies mentioned places by consulting existing gazetteers and existing placename lists. It also has the ability to inspect the context where place names appear in order to assert if they are geographical entities or not (i.e., Paris can be both the French capital or an individual's name).

With places being identified it is now necessary to locate them geographically by assigning coordinates to each of them. The TGN is used to assert to which real life entity a place name is referring to. This helps filter out cases of different places in the world with identical names. The opposite scenario of a place having multiple names can be filtered out as well by finding the main designation of the place in TGN. There may be cases where it is simply not possible to locate a specific place through the extracted name. In those cases, a default location is assigned that is usually the location of the largest or more often found place.

The final task is to model the boundary region corresponding to the initial vague place. First, it is necessary to calculate the spatial density of the previously extracted places that fall under boundaries of the vague region being determined. The density of a certain point *q* where *C* is a circle centered on *q* with radius *r* and *p* represents a relevant point, is given by:

$$\rho_q = \frac{\sum p_i \in C(q, r)}{\pi r^2}. \tag{3.7}$$

Equation 3.7 is a simple naive method which considers the influence of all points towards the density as equal. The influence of points is, however, better represented through a Kernel density estimation

(KDE) based around the distance of points to the center *q*. The closer to the center a point is the higher its weight value is. The application of this method requires the selection of the resolution to use and the value of the kernel radius *r* (in Jones' method the best radius is obtained through a trial and error exercise). The final result is a 2d grid cell, based on the chosen resolution, with each cell having a specific density value. A density threshold is defined (e.g., one point in grid cell) and the geographical representation of the vague region will correspond to the union of all the cells whose density value surpasses the chosen threshold. Jones spatial density based method provides results mostly in line with the regions the vague place designation usually refers to. A common problem, however, is that the achieved region tends to be larger than the real area associated to the vague place.

Another relevant work on defining vague places is a multi kernel learning approach proposed by Cunha and Martins [30]. The author's proposed method can be seen as an evolution of Jones method for defining vague regions since it follows the same overall template of finding associated points via web sources and then defining the vague region's boundary through them. However, besides minor changes in search methods and sources, the major changes probosed by the authors are related to the delineation methods used.

Cunha's region delineation method is based around one-class supervised learning on data collected from photo service sites (e.g., Flickr). The main goal is to, once again, use points, with descriptive features, that belong to the vague region in order to determine a boundary region corresponding to their contained area. However, the data on associated points may not include all the points that fall under the vague region's scope. Therefore, a supervised classification procedure modelled around the gathered points from available sources is used. This model is then applied to all possible points available which determines whether they are part of the vague region or not. Accuracy is given by the intersection of the region estimated by the model and the real life footprint of that vague region.

The supervised model of choice is a one class support vector-machine (SVM), widely used for data classification, whose main goal is to decide if the input data has the same properties of the training data or not. Data considered to not share the same properties are referred to as outliers. To accommodate outliers a parameter corresponding to the maximum number of outliers is defined. The similarity value of two data entries is given by a kernel function. Given two data examples *x* and *x'*, with a free parameter γ controlling the width, the kernel function *k* is defined as:

$$k(x,x') = \exp\left(-\gamma||x-x'||^2\right) \tag{3.8}$$

The final result of the model's classification is a weighted linear combination of all calculated kernel values plus a bias term β. Cunha's proposed method goes beyond this by using multiple kernels to determine the classification. Equation 3.9 represents the use of multiple kernels (in this case *M* number of kernels) to determine the classification of a target *x*.

$$f(x) = \left(\sum_{i=1}^{N} \alpha_i \times \sum_{j=1}^{M} \delta_j k(x_i, x) + \beta \right) \tag{3.9}$$

Where δ is the weight of each individual kernel method, α is the weight value for each training example and *N* being the total number of training examples.

Kernels methods are selected before the application of the model proposed by Cunha and Martins using a combination of Gaussian kernels with different parameters and kernel widths (γ). Overall, Cunha's approach of multiple kernels provides interesting results of relatively high accuracy. Despite this method still has some drawbacks mainly the fact it is highly dependant on the veracity of the training data used as well as the lack of proper metrics to determine the value of the region determined for other related fields.

3.3.3 Inferring Temporal Periods

A place can have multiple designations as well as different geographical footprints across many periods of history. Historical gazetteers deal with the representation of historical places by assigning a date or period to every feature of a place. However, comparable to the spatial representation of vague places, it may be hard to determine exactly when certain features of a place are valid or even exist [31]. The challenge then is how to correctly determine and represent the valid temporal period(s) of a feature. This task faces some complex issues ranging from vague temporal descriptions (e.g., around the early 1820's) to determining the transition between historical periods (e.g., end of middle ages).

There are many different proposals for effective temporal representation of events [32]. One of these is a proposal by Kaupinnen et. al [33] to use fuzzy sets to determine relevance of temporal intervals. Kauppinen uses the fuzzy set theory as the basis for modelling uncertain time ranges. Instead of simply considering an item in time to either be part of an interval or not, a membership value μ is attributed for each relevant item. This membership value varies between 0 and 1.

Kaupinnen defines a temporal interval as having four major components: *fuzzybegin*, *begin*, *end* and *fuzzyend*. The *fuzzybegin* and *fuzzyend* correspond to the earliest possible start of the interval and the latest possible ending of the interval, respectively. Meanwhile, the interval between *begin* and *end* is considered with certainty to be a part of the fuzzy temporal period, therefore all items within it have a membership value of 1. These values are represented on a 2d graph (Figure 3.4) with the X-axis corresponding to the time and the Y-axis being the membership value.

Through this type of representation it is also possible to determine the overlap value of two imprecise temporal periods even if in cases of fuzzy intervals on both sides. Another metric is related to the

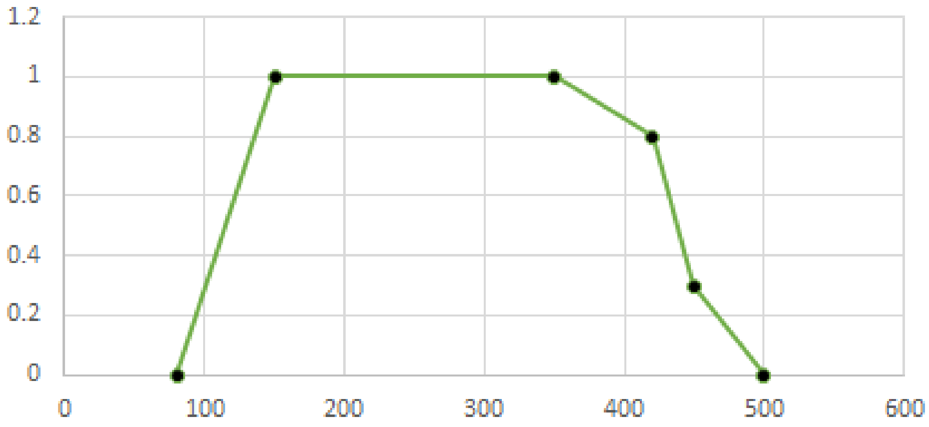


Figure 3.4: Example of a timespan representation.

closeness between two intervals, which is useful for intervals that don't overlap but are still close to each other in time. Finally, these overlap and closeness values are used to determine the relevance of a temporal period when compared to another.

Karl Grossner, the creator of the WHG, utilizes Kaupinnen's model of temporal representation as the basis for Topotime [5], a tool designed to situate complex and vaguely timed events or periods in a timeline. Topotime uses the same quadruple representation of fuzzy sets with additional options for the fuzzy sections representation (i.e., subintervals and curves). The membership value is also maintained, although in Topotime it is mostly referred to as likelihood instead.

Topotime's main feature is the ability to provide queries with a timespan, returning all the time periods where the query overlaps, ordered by the intersection area. The overlap value is the mean likelihood of the intersection. This can be of particular use when trying to determine the temporal location of a feature in the context of a gazetteer creation.

Other works for temporal representation in geographical information systems exist [34], however Karl's approach to defining vague temporal regions was considered to be the most accessible and comprehensive method for the context of this project proposal.

[5] http://dh.stanford.edu/topotime/docs/TemporalGeometry.pdf

4

Gazetteer Architecture and Development

Contents

Like established in the previous chapters, this thesis primarily concerns the creation of a historical gazetteer framework capable of handling many of the necessary data integration challenges that come with the development of this type of software. The first half of this section focuses on the describing the chosen approach and structure for the proposed historical gazetteer. This includes what related works were utilized and how they fit into the overall structure and functionalities of the gazetteer framework. Sources of data used to populate the gazetteer are also listed alongside a brief description of the type of data each of them contributed to the database.

Afterwards, the latter portion of this chapter shifts the focus into detailing the actual development process of the gazetteer's main components and functionality, according to the previously described architecture. This development process is itself divided into these respective main components of the gazetteer framework for a cleaner presentation.

4.1 Framework for a Digital Historical Gazetteer

4.1.1 Structure and Functionalities

The introductory chapter established the main focus of this thesis as the creation of a gazetteer framework capable of performing data integration tasks from different sources. The gazetteer also aims at providing that data to the users in a multitude of universal formats which can be exported from the gazetteer interface itself. Naturally, this type of framework required multiple different components working in parallel in order to accurately and effectively handle these different type of challenges. Therefore, in order to better organize the entire structure and its corresponding description, the entire gazetteer framework was divided into the following 3 main components: (i) the handling of data within the gazetteer's database; (ii) the user interface; and (iii) the delineation of regions for vague places within the gazetteer's database. Each of these components contains functionalities that fulfil one or more of the main objectives established at the start of this project. The development of these features made use of contributions from previously described related works as well as software and functionalities created during the development of this research project.

At the core of the historical gazetteer is the database storing place data and its schema format used which is paramount to the entire structure of the gazetteer. In this specific scenario, it was necessary to have a schema format capable of handling all the elements that compose the concept of a place which includes their geographical geometries, type and category. Figure 4.1 shows a visual representation of this concept of place alongside its temporal information. The ADL's format provided a great format to effectively handle this type of place data representation and, therefore, it was adopted as the base schema format for the gazetteer database. Therefore, all source data used were integrated into the ADL format which provided a stable basis from where to export place data into other more universal use

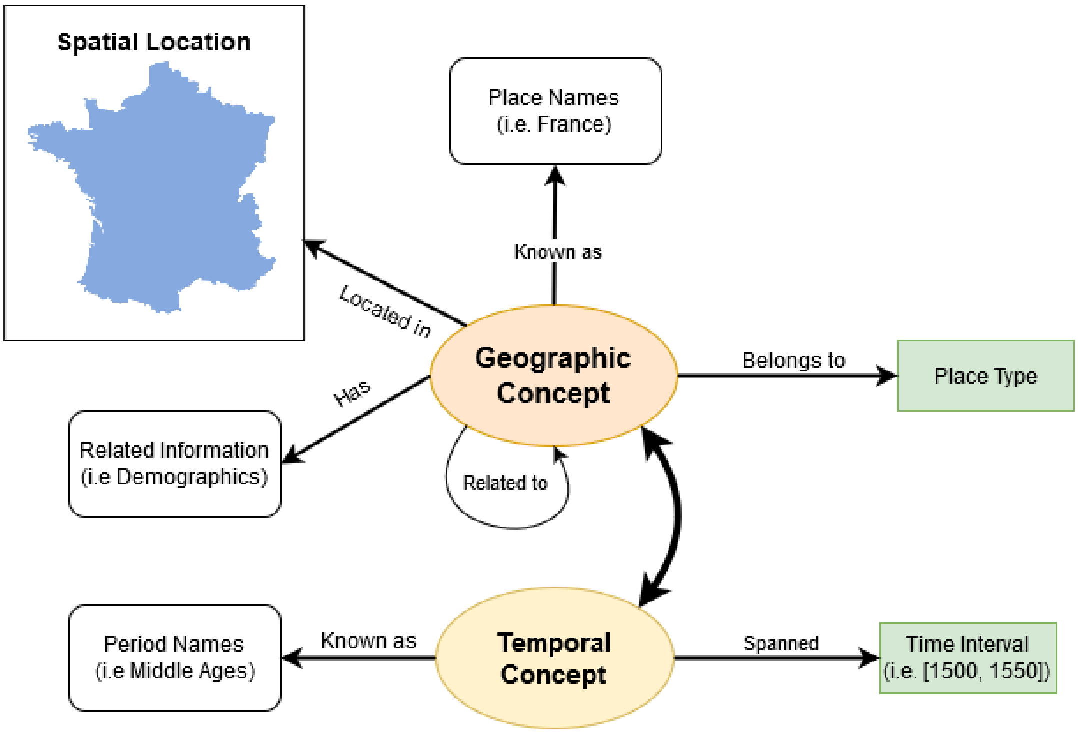


Figure 4.1: Concept and components of a place in the gazetteer.

friendly formats.

On this note, another important functionality of the proposed gazetteer framework is the ability to export its data in a variety of universal formats. Chapter 3 describes the LPF from the WHG which provides a good export format for any gazetteer project. While the LPF format is perhaps the best possible format in giving a complete representation for place data in the database, it is also possible to export data from the gazetteer into other formats, each of which providing different advantages. The first alternative option is to export data into a *CSV* file which provides a more compact and simplified information of the selected place data allowing the user to more easily organize and later use such data. It is also possible to export data into a shapefile (*SHP*), a more visual oriented format with the main focus on providing a visual representation of the spatial footprint for each place.

Following this, a reasonable amount of place data had to be imported into the gazetteer database in order to populate it and serve as a testing sample. This process, while essential in populating the gazetteer's database, required a considerable work in data integration tasks given the type of sources used which did not follow the same data formats adopted for the developed gazetteer. From this process it is relevant to highlight the importance of the data integration tools used. In particular, the detection of duplicates which is a core aspect of the integration process making use of some of the string matching techniques and geometry comparison methods presented in the previous chapters. Additionally, when dealing with vague places (e.g., lacking a spatial footprint) it was necessary to generate their geographical footprint based on their existing information (i.e. boundary box, related features and relevant points). Both of these alongside other relevant steps in the data import and integration process are described in more detail in Section 4.2.

Furthermore, there was still lacking a good interface to allow users to explore all these functionalities and stored data in an easy way. This led to the development of a web application with the purpose of handling all user interactions from simple place searching all the way to higher level data management features. The portion of the web application handling the searching of places and representation of their spatial footprints was heavily based around the Spelunker application from the Who's on First project. Therefore, it shares many of its characteristics including the layout design and the type of information being presented, offering an accessible and easy to use tool to visualize the data stored in the database. This web application also includes an advanced access level which allows the user to perform actual operations in the database such as altering tables or inserting new entries.

An initial prototype version of a gazetteer framework following these guidelines already existed with some baseline features implemented. This initial version contained the database based on the ADL format and a limited version of the web application still lacking many of the features previously described, in particular the visual search and representation of places. The main bulk of work done during this thesis consisted of improving this initial version of the gazetteer framework in order to fulfil the established goals of creating a gazetteer capable of dealing with the data integration and export challenges. The resulting framework is also flexible regarding future changes which in turn allows for many possible future improvements to be easily added either to the gazetteer itself or to its data management functionalities.

4.1.2 Datasets

Like previously stated, to test the functionalities of the developed gazetteer it was necessary to populate it with useful place data to serve as a sample. In particular, data from different sources to test the data integration process, with at least one source making use of temporal data to test the temporal data classification aspect of the gazetteer. Sources with missing features were also of value since they provided data for testing the polygon data generation tool.

Sources fulfilling the previous criteria included the GeoNames [35] geographical database, the TGN and the Digging into Early Colonial Mexico [1] (DECM) project. The datasets acquired from these sources provided enough data to comfortably operate the historical gazetteer and its corresponding features.

First, GeoNames is an open geographical database that covers all the countries in the world providing over 25 million different geographical names. All the gazetteer features are categorized into 9 unique classes and subsequently into one of the 645 available feature codes (essentially subdivisions of the

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

unique 9 classes). The GeoNames database is a valuable source since its geographical data is acquired through the integration of data from multiple other sources. The data in GeoNames is made available through a variety of web services which are updated daily guaranteeing the most accurate data at the time of importing.

The TGN, previously presented in Section 3.1, was also used as a source for place data. As mentioned before, TGN provides a vast number of place names and alternative names from multiple languages and historical periods. Furthermore, the data in the TGN is made available as Linked Open Data making it an easy process to extract the relevant data and integrate it into the proposed gazetteer.

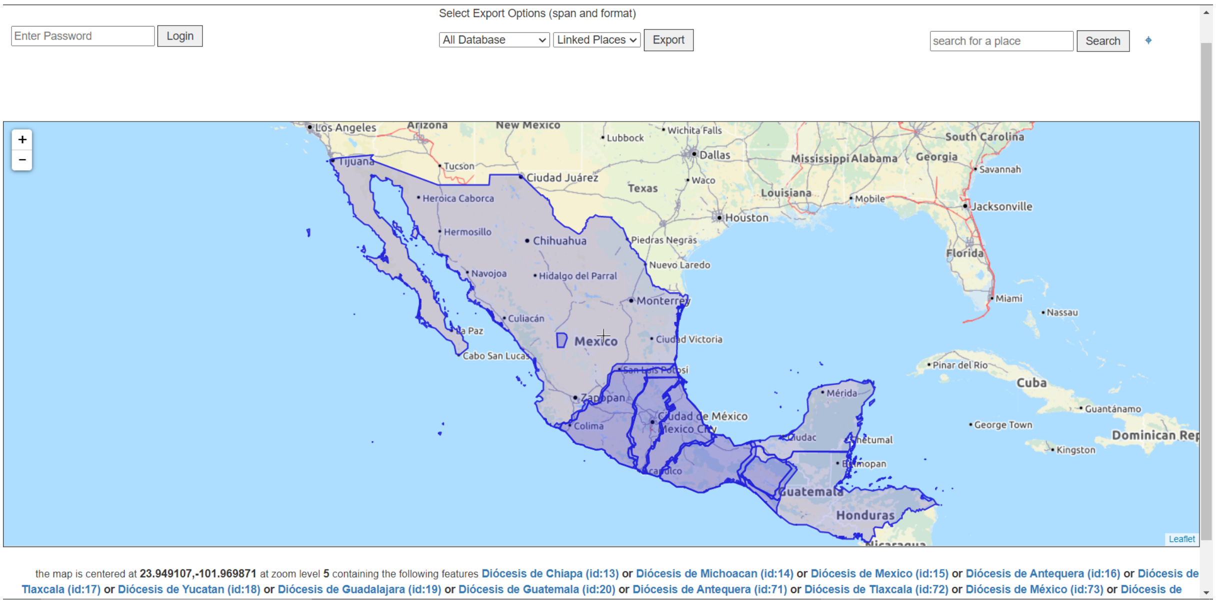
Finally, data from the DECM project was another source used to populate the gazetteer. The DECM project developed a digital approach to explore data from the early colonial Mexico period. This data has been collected from a collection of Spanish official documents from the era present in the corpus *Relaciones Geográficas de la Nueva España*. This source is considered to be one of the most important for the early colonial America period, although its massive scale and complexity made it impossible to access in a effective way. However, through the DECM project it was possible to access the data which proved to be valuable to test the temporal features of the gazetteer framework given its historical coverage.

Given the historical nature of the gazetteer, it needed to take into account the temporal aspect of the imported place data such as identifying to which historical period a place in the database belongs to. The PeriodO [2] project was the best source of data found to handle this temporal aspect. Providing a database linking scholarly time periods to accurate time intervals (e.g., ”3rd Millenium BCE” corresponds to the year interval [-3000,-2000]), the data from PeriodO was a quick and easy method for determining a vague time period, avoiding the need for more complex methods in order to place a vague period in the historical timeline.

4.2 Development Process and Components of the Gazetteer

This section details the development process of the necessary parts for the proposed historical gazetteer including the supporting integration tools and gives a general description of the web application functionalities and layout. The description of the entire process was divided into three main parts, although they were developed in parallel. First, a detailed description of the web application and its features including both the pre-existing ones in the original repository as well as the ones added during this development. The second part details the process of importing and integrating place data into the gazetteer database such as handling duplicates and creating a network of relations between the remaining places. Finally, the last part describes the generation of spatial geographical data for vague places with a spatial footprint

[2] http://perio.do/



the map is centered at **23.949107,-101.969871** at zoom level **5** containing the following features **Diócesis de Chiapa (id:13)** or **Diócesis de Michoacan (id:14)** or **Diócesis de Mexico (id:15)** or **Diócesis de Antequera (id:16)** or **Diócesis de Tlaxcala (id:17)** or **Diócesis de Yucatan (id:18)** or **Diócesis de Guadalajara (id:19)** or **Diócesis de Guatemala (id:20)** or **Diócesis de Antequera (id:71)** or **Diócesis de Tlaxcala (id:72)** or **Diócesis de México (id:73)** or **Diócesis de**

Figure 4.2: Result page of the search query ”Diócesis”.

lacking in the database.

4.2.1 Web Application

The web application acts as the main interface of the entire gazetteer framework with all user interaction being done through it. The application itself can be divided into a basic access level for basic searching of place information and an advanced access level that allows direct interaction with the database schema and records. The basic level was entirely added during this research project while the advanced level was already part of the preexisting gazetteer framework present in the initial repository. Taking this logical division into account both parts will be described separately in the remainder of this subsection.

On a technical level, the application was built using Flask [3], a highly flexible Python based web application framework. It provided an easy and effective way to create this user interface and all the necessary database connections and operations it needed to fulfil the proposed functionalities.

Upon opening the application the user is presented with the basic access level interface which consists of a top section and an interactive world map. The first tool in the top section is the login tool through which the user can authenticate himself and gain access to the advanced level. At the middle of the top section is the data export tool where the user can export either the entire database or just a smaller selection of places displayed in the map bellow in any of the available formats: LPF, CSV or a Shapefile. The resulting file is automatically downloaded from the browser into the user's local storage.

At the top right of the interface lies the search bar which allows the user to search for a place through

[3] https://palletsprojects.com/p/flask/

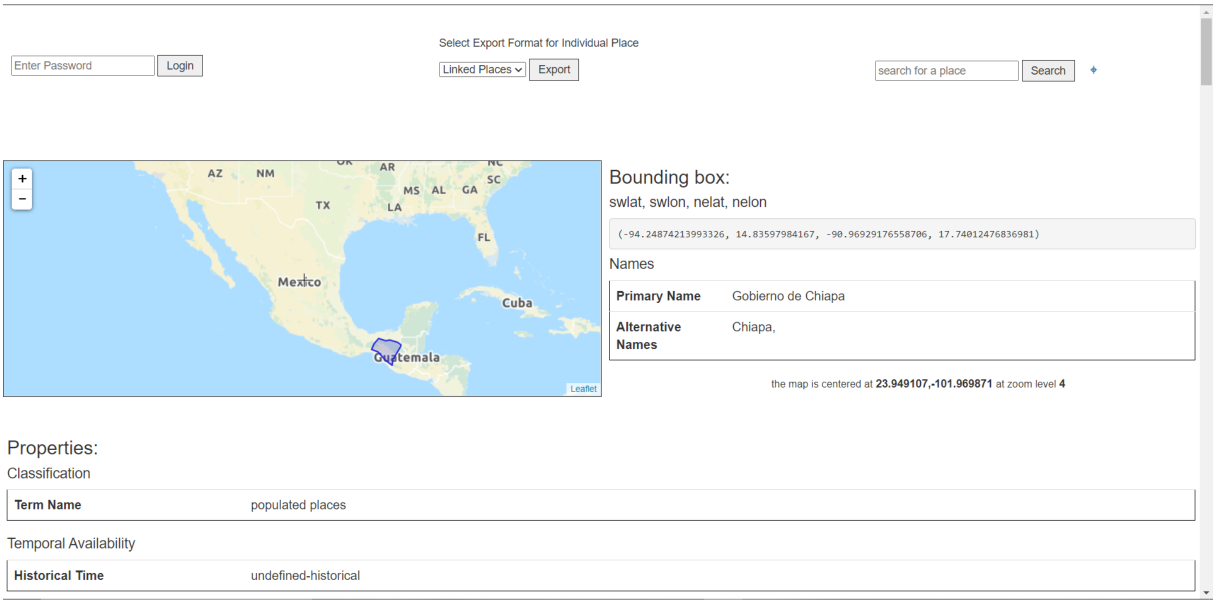
one of its known designations. The search tool dynamically provides auto-complete suggestions by gathering possible places from the gazetteer database according to the current input as it is being written. When the search query is submitted the application retrieves all possible place entries in the gazetteer corresponding to the search term being searched (e.g., searching for the term ”Yukatan” provides all the results containing that term in its designation no matter the position, such as ”Provincia de Yukatan” or just ”Yukatan”). The places from the results are then presented on the map with their corresponding name and identifier in the database (the latter of which helps distinguish between similar places).

Given that the data used for this project is mostly centered around the area of central America the map is initially centered on that area. The major goal of the interactive map is to present the geometries of places that fall under a search query provided by the user. Figure 4.2 shows the result of the search query ”Diócesis” which consists of all the places in the database with that term in their designation. It is also worth to note that in the case of overlapping spatial footprints (such as in the central area of Figure 4.2) the user can click on any polygon which prompts a toolbox identifying the place that corresponds to that geometry.

A place can have either a polygon or a point type of spatial footprint so for ease of representation they are colored blue and red in the map, respectively. Furthermore, a list of all the features presented on the map is available at the lower section of the map, and selecting one of them directs the user to a page with detailed information on the selected place. This list of information starts with the general bounding box of the place followed by its properties which includes: primary name, alternative names, placetype classification, temporal information, geometry (coordinates, area and map representation) and the sources from where the information was gathered from. Lastly, a list of related features is also available at the bottom of the page, with each relation including the name of the related place and the type the relation. Selecting a row from the related places list directs the user to the related place's individual information page. It is also important to highlight that the top section of the initial page remains unchanged allowing the user to login, export the data of the page's highlighted place or search for a new place which redirects the user back to the initial search page with the results.

After logging in to the advanced access level the user is presented with the main page of this section which allows the direct interaction with the gazetteer's database. This page lists some general information about the gazetteer schema and files such as their names, number of tables/views and dates of modification/creation. At the top right right corner is the create table input box which provides an easy way for the user to create a new table by just inserting the name for the new table. To the left are all the current table names available in the gazetteer schema and by clicking one of them the interface for that table is opened.

The individual page of each table is composed of 3 tabs: Structure, Content and Query. The Structure tab shows the more general information of the overall table and its attributes. This includes a SQL



Bounding box:

swlat, swlon, nelat, nelon

```
(-94.24874213993326, 14.83597984167, -90.96929176558706, 17.74012476836981)
```

Names

Primary Name	Gobierno de Chiapa
Alternative Names	Chiapa,

the map is centered at **23.949107,-101.969871** at zoom level **4**

Properties:

Classification

Term Name	populated places

Temporal Availability

Historical Time	undefined-historical

Figure 4.3: Page for individual place information

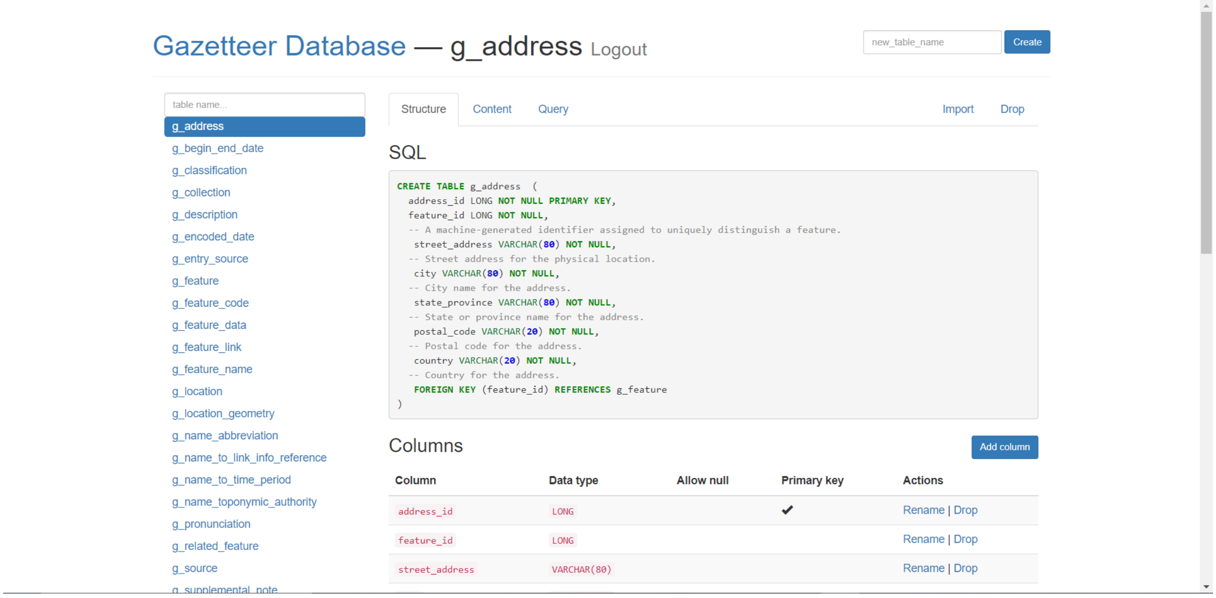
query to create the table with its current schema alongside with relevant properties such as column descriptions, foreign keys and indexes. The Content tab shows a list of all the records present in the table. Finally, the Query tab allows the user to execute SQL queries to that table as well as to export the results of the query into one of the available exports formats. In all tabs, it is also possible to import more data into the table or to drop the entire table.

With the functionalities of both levels, the web application fulfils its intended role of acting as a gateway between the user and the database while allowing the interaction and visualization of the gazetteer's data and information. Furthermore, the data export tool in particular allows this framework to act as an universal data provider by exporting its data in easy to use formats for other projects.

4.2.2 Data Integration Process

The data import and integration processes are at the core of populating the gazetteer database. Formerly, the DECM project which compiles place data from colonial central America was established as this project's primary source of place data. However, before importing data from the DECM project it was necessary to do a preliminary import of mostly basic metadata. In particular, a list of placetypes from the TGN project for place classification alongside the historical intervals and their respective denominations from the Period0 project for temporal place classification.

After establishing the basis for the database with these initial imports, the next step was to import and integrate actual place data from the primary sources of choice. The first import of place data came from the GeoNames database with some places relevant to the location of central America being imported into



```
CREATE TABLE g_address  (
  address_id LONG NOT NULL PRIMARY KEY,
  feature_id LONG NOT NULL,
  -- A machine-generated identifier assigned to uniquely distinguish a feature.
   street_address VARCHAR(80) NOT NULL,
  -- Street address for the physical location.
   city VARCHAR(80) NOT NULL,
  -- City name for the address.
   state_province VARCHAR(80) NOT NULL,
  -- State or province name for the address.
   postal_code VARCHAR(20) NOT NULL,
  -- Postal code for the address.
   country VARCHAR(20) NOT NULL,
  -- Country for the address.
   FOREIGN KEY (feature_id) REFERENCES g_feature
)
```

Column	Data type	Allow null	Primary key	Actions
address_id	LONG		✓	Rename | Drop
feature_id	LONG			Rename | Drop
street_address	VARCHAR(80)			Rename | Drop

Figure 4.4: Component to interact with the database schema and records.

the Gazetteer. Afterwards, it was the main import of place data from the DECM project into the gazetteer database which was divided into multiple files with some of them containing large administrative areas as well as more specific ones. Each file contained large lists of places and their relevant information within the covered area. Nevertheless, the import process was similar for both types which started by first inserting all the existing information (i.e., names, type, location, etc) into the database followed by creating possible relations with other places either described by the files itself on through association by filename or location.

While the previous explained import process was responsible for the majority of the entries in the database, there were still necessary further integration steps in order to maximize the information retrieved from the imported data in the gazetteer database. During the data import process it was noted that some source files often provided lackluster or vague information regarding how the different places related and affected each other. To counter this issue, an additional integration step was created which analyzed all the geometries present in the gazetteer database and created relationships between places based on the spatial analysis and comparison of their respective spatial footprints. The places could then be labelled as adjacent, identical, overlap among other types of relations. The resulting chain of relations created an all encompassing hierarchy that better contextualized every place within the gazetteer.

The next step involved establishing the temporal information of all the places previously imported into the gazetteer database. In the historical gazetteer all places contain a classification which describes the time period when that place was active. However, the sources used during the main import process didn't include temporal information on the places listed which led to all the places in the database having their primary temporal period labeled as unknown. In order to make use of the historical aspect of the gazetteer, historical documents from the colonial era in question were gathered and used to identify which places correspond to that era in time. Initially, the process primarily consisted of simply searching for the name of places in the large historical documents. If a match was found the place in question as belonging to the ”mesoamerican colonial era”. This initial approach proved lackluster so more accurate sources that better described the places present in the documents were utilized. These sources consisted of a set of files with the list of placenames present in each of the historical documents which removed the need to analyse the entire text of the documents. In the end, the process to determine a place's temporal information consisted on comparing the places mentioned in the aforementioned historical documents with places already present in the gazetteer database. In the case of a match, the place data in the gazetteer was updated as an historical place along with the document in question being labeled as the source for that particular change.

Finally, after processing both spatial and temporal place data, the last step involved the detection and handling of duplicates still remaining in the database. First, all places with defined locations were analysed among each other to determine which pairs of places shared a similar location using a distance of less than 500 meters between the centroids of both places as a threshold in the case of points. In the case of more complex spatial footprints, such as polygons, both regions were analyzed and in case they were considered identical they were also marked as a possible duplicate. This process is done through the python library Geod [4] which contains several methods for geometric computation, including calculating the distance between two locations. Next the names of the possible duplicate places were compared through a Jaro-Wrinkler based string similarity algorithm with a threshold similarity of 0.9 (i.e., above this similarity value they are considered a duplicate). In the case of a positive match in both of these conditions, the pair of places were considered duplicates and marked as such with a new relation of equivalency between both places in the database. This solution leaves the possibility for future updates where the duplicate pairs can have one of the places removed, both places merged into a single entity, or only one of them considered as the main place and visible in the web application.

4.2.3 Polygon Data Generation

Last in the list of the main functionalities of the gazetteer framework is the polygon generation tool which generates spatial footprints for vague places lacking such feature. It was previously established the existence of places without a corresponding spatial footprint stored in the database either from missing information from the sources used or due to the vague nature of the place in question. A tool script that generates polygons for this type of entries was developed in order to provide a better spatial context for these places and maximize the amount of information available in the gazetteer as a whole. The tool used a list of relevant points from a set of places (usually of the same type for more consistent results)

[4] https://pyproj4.github.io/pyproj/stable/api/geod.html

and created a raster file based on the Voronoi distance algorithm between each unit of area and all the reference points. This raster file is essentially a bitmap divided into multiple units of area each with a value associated to it corresponding to one of the places used as input. The values of the resulting raster file are then extracted and used to create a shapefile with multiple polygons, each corresponding to one set of values of the original raster file. From this initial diagram, which effectively represented the area of influence of each vague place, a polygon was attributed to the vague place whose reference points were responsible for generating the area.

Despite this, a simple Voronoi diagram proved to not be enough to accurately determine a place's spatial footprint. Therefore, weights were added to each point based on the population density from the source Gridded Population of the World [5] project. Reference points in higher population areas were given a higher weight in the distance based Voronoi algorithm which calculated the values of each unit of area from the raster file. This in turn resulted in the creation of a diagram with a more accurate representation of the real spatial footprints for the vague places used as input since more context was added to the area distribution method instead of a simple blind attribution based on the distance alone.

Afterwards, the polygons were extracted and simplified into a single shapefile and intersected with polygon of central America to crop the areas falling in the sea. A single polygon was attributed to each of the places whose reference points contributed to generate the final shapefile. This attribution was based on checking which polygon contained most of the reference points provided by each place and giving out a score based on it. The polygon with the highest score was considered to be the corresponding representation of that specific place and inserted into the database as such. The final result provided a much more comprehensive database since the places previously lacking their spatial data could now have a relatively accurate geographical representation whose spatial relations with other places in the database provided better context for that place within the gazetteer's overall hierarchy.

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

5

Results and Analysis

Contents

Chapter 5 presents the results of the developed historical gazetteer and their evaluation. The first Section 5.1 establishes a base method upon which this gazetteer was evaluated in a general way. Afterwards, Section 5.2 presents data and statistics of the gazetteer's database allowing an analysis of its final state after executing the data import and integration processes. Meanwhile, Section 5.3 focuses on the evaluation of the integration tools used in scenarios created specifically for that purpose, in particular, the duplicate detection tool and the polygon generation tool.

5.1 Methodology and Evaluation Metrics

Evaluating a gazetteer is not a straight forward task given the lack of universal metrics to determine its overall quality and accuracy. Even with these limitations there was still the necessity of evaluation and comparing gazetteer projects in order to identify the contributions of the different types of other gazetteer projects. Some of these projects attempted to establish an effective way to evaluate the final state of a gazetteer and provided useful methods for evaluating this thesis' gazetteer and its components. Among these is the work of Acheson [5] whose goal was to analyse and compare the world spanning gazetteers of the previously mentioned GeoNames and TGN projects. Acheson's method made use of a definition that ranks a gazetteer's individual quality according to the following criteria:

1. **Availability:** The availability level of the gazetteer;
2. **Scope:** Overall coverage of the gazetteer (i.e. region, country, world);
3. **Completeness:** Degree of gazetteer coverage;
4. **Currency:** Time interval for changes incorporated into the gazetteer;
5. **Precision:** Reliability of places in the gazetteer;
6. **Granularity:** Size of features;
7. **Balance:** Uniform degree of currency, accuracy and granularity;
8. **Richness of annotation:** Quality and number of descriptions within the gazetteer;
9. **Lineage:** Sources used for places in the gazetteer.

Additionally, Acheson's comparison method also utilized a wide variety of map representations (i.e. density maps or placetype maps) to better visualize and evaluate data accuracy and completeness. The type of maps varied in detail and scope but shared the primary objective of asserting if the place data was in line with real data on the same places from other reliable sources. Although this evaluation and analysis method was used to compare 2 different gazetteers, it also proved quite useful for analysing a single gazetteer framework by determining if the distribution of the data made sense contextually (e.g., terrain entity locations corresponded to topographic reality).

Taking this into consideration, Acheson's analysis and evaluation served as the basis for the evaluation of the historical gazetteer regarding the final state of its place data. Initially, the state of the gazetteer data was evaluated according to each of the previously listed Acheson's criteria. Some criteria lacked critical information to be labelled with certainty so adjustments had to be made as necessary. Following this initial preset evaluation, all kinds of statistics and numbers within the gazetteer were listed in order to provide an overview of the general state of the data in more detail. This list included the total number of places, designations, placetypes and places with spatial representation as well as their respective geometry type. Additionally, data originated from the integration process was also detailed which included stats such as the number of duplicates, the number of relations with type distribution, the number of detected historical features and the number of generated spatial footprints.

In parallel to this overall analysis, multiple maps related to place distribution were created in order to easily visualize the distribution and location of the places inside the gazetteer. These maps differ from the ones used by Acheson's to better highlight the attributes relevant to the context surrounding this thesis' gazetteer. Therefore, the following maps detailing the distribution of different places were created:

- Overall point density map;
- Distribution of relevant places by types;
- Map comparing historical and non-historical places;
- Compare the terrain map of the area with the distribution of terrain type places in the gazetteer.

These data, statistics and different types of maps ultimately painted a comprehensive picture of the overall quality and accuracy of the developed gazetteer. This allowed us to determine whether or not the initial goals defined for the creation of this gazetteer were met or not.

However, the evaluation of this project still required a more detailed analysis on the data integration process. In particular, the main tools whose evaluation was deemed as relevant were the duplicate detection tool and the polygon generation tool for vague places. Unlike the previous general evaluation process, the accuracy of both of these tools was determined by testing them in datasets or scenarios that allowed us to evaluate their respective overall performance and identify any issues or limitations.

In order to evaluate the performance of duplicate detection tool it was necessary to test it in a scenario with a large number of duplicates. For the purpose of this evaluation, a dataset from the GeoNames [35] project containing a large number of placenames pairs with the corresponding classification annotations was used as a test scenario. However, it was still considered important to check the performance of the tool on the gazetteer itself. Thus, a sample of 50 pairs of duplicates detected during the data integration process was randomly selected and manually evaluated, which provided an insight on the performance of the process.

Meanwhile, the evaluation of the polygon generation tool revolved around attempting to recreate regions that were already stored in the gazetteer database with the correct shape. In particular, reference points were retrieved from a set of places sharing the same placetype which already had polygon representations in the database and used as input to recreate those same original representations. Afterwards, the generated set of polygons were compared with the original ones in order to determine how accurate the recreation attempt was. This approach allowed us to evaluate the tool while also allowing the option highlight its strengths and weaknesses by presenting specific scenarios that highlight each of them.

5.2 A Case Study with Early Colonial Mexico Data

5.2.1 Gazetteer Quality Evaluation

According to the aforementioned evaluation method, the state of this thesis' gazetteer was analysed using Acheson's criteria for gazetteer quality presented in the previous section. The following Table 5.1 presents the defined evaluation.

Availability of the gazetteer was labelled as *Free* since the gazetteer is available without restrictions outside of the advanced access level which requires an authentication. Furthermore, the region of Central America was considered the scope of the gazetteer given that the DECM was the primary source of place data for populating the gazetteer database.

Meanwhile, the lack of a guideline gazetteer from this area and historical period makes the task of determining the completeness of the gazetteer a difficult one. It is probably safe to assume that the gazetteer is not a 100% complete representation of all places under its chosen scope. At the same

Criteria	Thesis Gazetteer
Availability	Free (data management requires authorization)
Scope	Central America Region (focus on Mexico)
Completeness	Unknown
Currency	Not determined
Precision	Approximate
Granularity	Low to Fine
Balance	Consistent
Richness of annotation	Medium
Lineage	Various Sources

Table 5.1: Thesis gazetteer classification according to Acheson's Criteria

time, the current gazetteer still provides a considerable number of places and information which provide a completeness level above 0%. As such, it was unclear the level of completeness of the final gazetteer and that criteria left as *Unknown* at the time of evaluation.

Given the nature of this project, the regularity and interval between future updates on data and functionalities has not been determined. This meant that the gazetteer's currency criteria had to be categorized as *Not determined* since there was no road-map for future updates to the gazetteer framework. Additionally, the level of precision of the places present in the database are inherently tied to the reliability of the sources utilized. This reliability is further explored in Section 5.2 where various maps showcase the logical distribution of places in the gazetteer database. These results pointed towards a highly accurate and precise level for the data present in the gazetteer which is inline with the sources of data used. However, the same level of precision could not be guaranteed to places whose spatial location had to be generated through the polygon generation tool. Therefore, the overall precision of the gazetteer was considered as *Approximate* being mostly brought down by the less accurate regions generated through the polygon generation tool which lacked the level of reliability of the external sources for place data.

The granularity of places in the gazetteer varies from very low to very fine precise locations which in practice means that places range from large administrative areas all the way down to specific small towns and villages. There is also the extensive hierarchy of relations between the places connecting the largest state with the smallest place. This hierarchy as well as the focused coverage area allowed the gazetteer to achieve a consistent level of balance across its overall scope. Making use of only a small number of sources also contribute to this fact since a larger number of sources usually contributes to larger differences in levels of granularity and detail.

Finally, a significant number of annotations and information is also available in the gazetteer although not to a very extensive and exhaustive degree. This area has a lot of room for improvement as the database already supports a large number of annotations across all the elements that compose a single place entity. A possible improvement for the future would be to add more detailed annotations and descriptions for places already in the database. The last criteria is self explanatory with various sources for place data being used to populate the gazetteer database.

5.2.2 Data Statistics and Correlation with External Sources

Besides the general quality evaluation of the gazetteer it was also relevant to consider the data and statistics of the database. These allowed for a more detailed evaluation than the previous approach based on Acheson's methodology. After the import and integration of place data, the final historical gazetteer contained a total of 47305 features (individual places) in its database. In terms of place names or designations the gazetteer contained 56118 place designations with only 41205 being actual

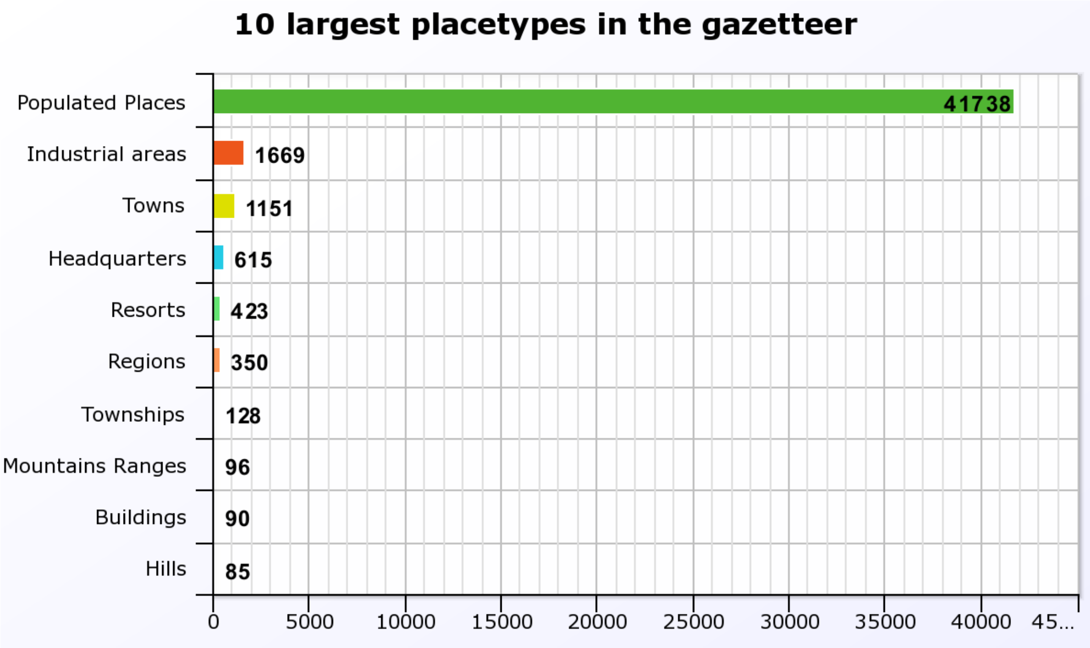


Figure 5.1: Number of Placetypes in the Gazetteer

distinct names. In particular, the maximum number of designations for a single place was 16 with the average across the entire gazetteer sitting at 1.19131, meaning the vast majority of places only have a single designation attributed. This difference between the number of distinct names and the total numbers of place designations can be caused by multiple uses of the same designation across multiple places which itself is a sign of some possible duplicate entries in the database. On this note, during the integration process a total of 5944 features were labeled as duplicates (i.e., they had a similar name and spatial footprint) which likely also contributed to the disparity between total and distinct placenames in the database.

Each place in the gazetteer database also has a specific type classification attributed to it upon being imported into the database. As such, looking at the respective data regarding placetypes showcased that the classification of places in the gazetteer was distributed across 66 distinct types. Figure 5.1 provides a visual representation of the numerical distribution of the 10 largest placetypes in the gazetteer. The first aspect to highlight is the large portion of places being labeled as *Populated Places*. There were two possible identified reasons that could explain such a large difference between the top placetype compared to the remaining placetypes. The first, and likely the most influential, is the fact that the term *Populated Places* is a common staple term widely used for generic places that lack a more specific category. Additionally, the term was also chosen as the default type to be used in cases where the placetype was not defined in the sources used, which further inflated its numbers across the database. Given the number of places lacking a type in the sources used, or just having a generic classification, the uneven distribution of placetypes was an expected outcome, given the reasons already presented.

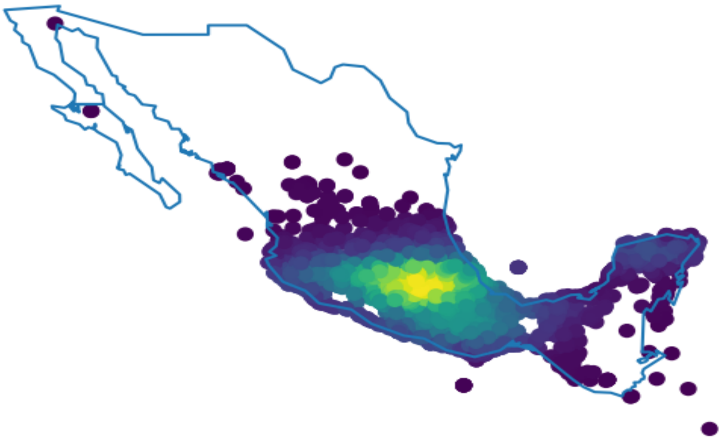


Figure 5.2: Density map of places in the gazetteer

Despite this unbalanced distribution, the remaining placetypes include terrain, buildings and even industrial areas which provided a varied selection of placetypes for the covered area. These remaining types collectively correspond to around 6000 places, still representing a significant number for the area corresponding to the early colonial Mexico. Furthermore, terrain related types (i.e. rivers and hills) could be used as a way to determine the accuracy of the data within the gazetteer by comparing their location with other sources.

Lastly, the remaining element that composes a place in the gazetteer is its spatial footprint. The geographical distribution of the features stored in the gazetteer provided a useful method to visually identify the gazetteer's coverage and completeness. The gazetteer contained 7956 imported geometries which were themselves composed by 7480 points, 368 polygons and 107 multipolygons. Furthermore, the polygon generation tool created 1748 new polygon geometries for features without geographic representation which brought the number of geographic footprints stored in the database to a total of 9704.

Figure 5.2 provides a density based map from all places with defined locations stored in the gazetteer database. As shown, the gazetteer provides an extensive coverage of the central Mexican area, in particular the area of interior Mexico, a major urban population center. The covered area and the concentration of points are consistent with the context of the sources utilized and the modern population distribution of Mexico.

It was also relevant to check if the placetype classifications for places in the gazetteer made sense in context and when compared with other sources. Figure 5.3 contains the distribution of some specific placetypes whose location could be used to determine the accuracy of the data (e.g., buildings located

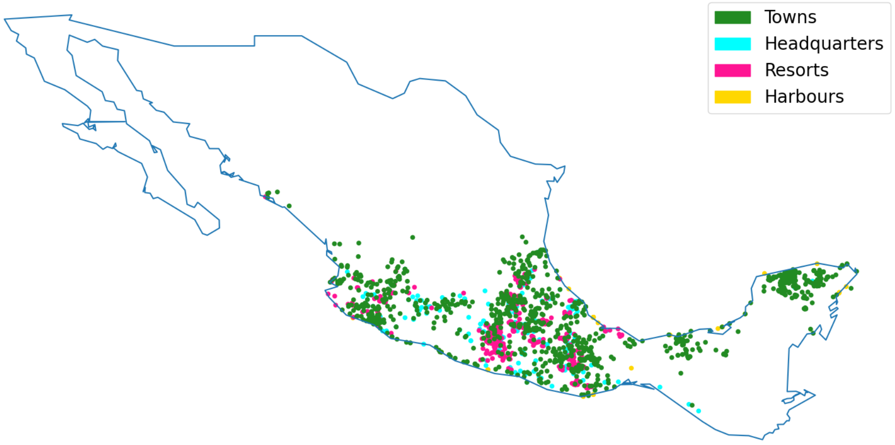


Figure 5.3: Distribution of the relevant placetypes in the gazetteer.

in population centers and ports located near the coast). This example shows that towns are overall more dispersed throughout the covered region while resorts (type that contains hotels and other touristic facilities) are more densely concentrated in the urban areas of central Mexico with some coastal outliers. Another thing to note are places with the type harbour largely located near the coast which is inline with their real function and use. Finally, headquarters (which in this context refers to municipal headquarters) are mostly dispersed across the entire covered regions without a clear high density area. This uniform distribution of headquarters is perfectly in line with the nature of the municipal division at the country level. Overall, the information provided by the location of places based on their type points towards a logical location and distribution given their surrounding context and main use or function.

Places whose type falls under terrain related categories (i.e., mountains and rivers) also provide an useful insight into the accuracy of the data stored in the gazetteer. It is possible to determine the classification accuracy of the type of these places by comparing their locations with real topographic maps of the covered area. Figure 5.4 contains the most relevant terrain type places with spatial representations in the database.

Figure 5.4 shows a high concentration of hills and mountains in the central interior. In particular, it can be seen a mountain range that spans vertically into the central interior region of Mexico. In this interior area the elevation levels are more subdued given the increase in number of hills appearing. Meanwhile, the rivers are located more on coastal areas with the eastern coast area containing a considerable number of rivers side by side.

Comparing this distribution with the real life topographic map of the area shown in Figure 5.5, it

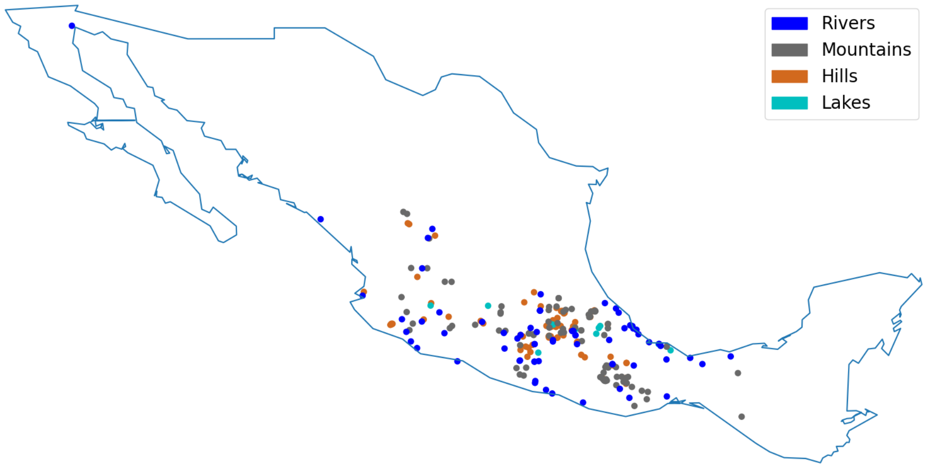


Figure 5.4: Distribution of the most relevant terrain types in the gazetteer.

is possible to see that the previously mentioned mountain ranges correlate with the topography of the central area of Mexico. The line of mountain ranges stretching from the southern area all the way up to the central interior area is present alongside the decrease of the levels of elevation towards that central interior area. Additionally, the large numbers of rivers ending in the eastern coastal area also correspond to a high degree of accuracy with the position of the places with the type river present in the gazetteer[1]. This similarity between the terrain information on the gazetteer and the real topographic map of Mexico is a sign favouring the accuracy of the gazetteer's place data regarding both type classification and their respective location.

This being an historical gazetteer, the exploration of the temporal aspect of the place data in the database was also a relevant part of this evaluation process. Like already described, the integration process identified a total of 2580 places as being from the 16th century or early colonial period. Figure 5.6 shows the distribution of non-historical and historical places with known locations in the gazetteer.

Overall, both historical and non-historical places share the same general distribution with a main focus on the more populated area of central Mexico and another concentration of points in the north of the Yucatan peninsula. Although there isn't a particular pattern to the historical places identified, the final number of historical features still provides a reasonably sized sample which allows for a comprehensive view of the status of the covered area during that particular historical period.

After analysing the data of all the different components of individual places, all that remains is to analyse the interconnected relations between them in the gazetteer. This network of relations creates

[1] https://geo-mexico.com/?p=9117

Figure 5.5: Topographic map of Mexico from Geology[2]

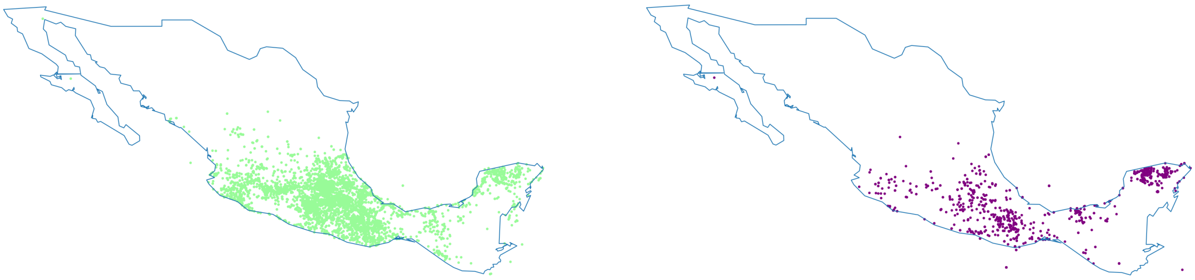


Figure 5.6: Distribution of non-historical places (left) and historical places (right) in the gazetteer

the contextual base of the gazetteer and provides information on how each place entity is related to the remaining entries in the gazetteer, and how it sits in the overall hierarchy of places. Relations vary between location based relations and context level relations which require further information besides the physical location of the places involved. At the time of writing, the gazetteer contained a total of 228914 relations between different places with an average of 26 relations per place. The maximum number of relations for a single place reached the number of 7294 corresponding to an administrative region with a high coverage area. Table 5.2 shows the types of relations that exist in the gazetteer with the respective number of entries for each of them. Meanwhile, Figure 5.7 presents this same information in a pie chart providing a better visualization of the distribution of all the relation types.

The vast majority of relations in Table 5.2, corresponding to 90%, are related to places which contain

[2] https://geology.com/world/mexico-satellite-image.shtml

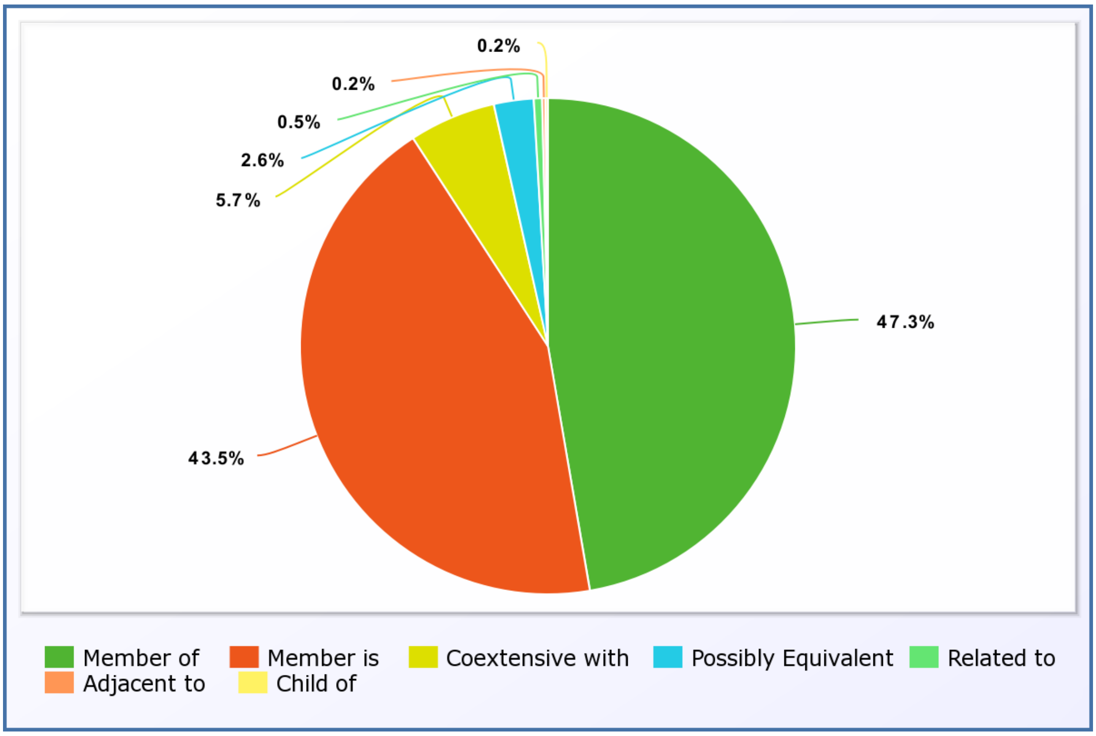


Figure 5.7: Distribution of the different types of relations.

or are contained within another place. This number shows the existence of places in the gazetteer with spatial footprints covering large areas of land such as states, administrative states or dioceses. Other smaller sized places are inevitably contained within some of these large areas which in turn represent most of the relations of this gazetteer. From the remaining 10% it is relevant to highlight the *Possibly Equivalent* relationship which was attributed to places considered duplicates during the data integration process with the total number of real duplicates being half of that, given that both places have the same relation type in the database. Lastly, a small number of place relations fall under the type *Related to*, which essentially covers relations based around non-location context such as historical or type related context.

Relation Type	Number in Database
Member of (contained inside another place)	108268
Member is (place is contained within it)	99674
Coextensive with (intersection between 2 places)	12950
Possibly Equivalent to (duplicates or very similar places)	5944
Related to (unknown type of relation)	1243
Adjacent to	458
Child of	377

Table 5.2: Number of relations in the gazetteer by type.

5.3 Integration Tools Evaluation

This section evaluates the overall accuracy and quality of the tools combined to perform the data integration process. In particular, the tools evaluated are the duplicate detection tool and the polygon data generation for vague places. Both of these tools were be tested on specific scenarios and datasets, and their results compared to correct duplicates and regions, respectively, in order to determine their overall accuracy.

5.3.1 Duplicate Detection Results

As described previously, the duplicate detector evaluation consisted on determining the performance of the tool when applied to a dataset containing a large number of placename pairs from the GeoNames [35] project, with the corresponding classification of each pair on whether they were duplicates or not. This allowed to test this tool across a large data sample which in turn gave a more reliable basis to analyse the results. The results from running the duplicate detection mechanism across the GeoNames dataset are depicted in Table 5.3.

Firstly, the chosen dataset proves to have a credible sample size with a total around 160000 pairs of names. In terms of results, the duplicate detector managed to correctly detect 341122 duplicates and classified another 824298 as not duplicates, a total of 70.5% accuracy. A more detailed look shows a sensitivity value of 93.77% which means 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% representing the main factor bringing down the overall duplicate detection accuracy.

The results of this test point towards a bias for limiting the number of pairs being wrongly considered as duplicates 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 of them removed in the future, this could result in a large number of non-duplicate places being wrongly removed

	Numbers/Results
Total Entries	1652962
True Positives	341122
False Positives	54753
True Negatives	824298
False Negatives	432789
Accuracy	70.50%
Sensitivity	93.77%
Specificity	44.07%

Table 5.3: Numbers and results for the duplicate detection in the GeoNames dataset

which essentially means the loss of important distinct places.

It is also relevant to determine the performance of the duplicate detection tool in the actual gazetteer itself. However, unlike the previous GeoNames dataset, there is no way to determine if a duplicate is correctly classified as such outside of manual checking. This is of course not realistically feasible given the current numbers of identified duplicates as 2972. The compromise found to handle this issue was to randomly select 50 duplicates from the database and manually analyse if their classification as duplicate or not was correct, not correct or otherwise unknown. This provided a small yet useful insight into the accuracy of the duplicate detection tool during the data integration process of the gazetteer.

After running the query to retrieve the aforementioned 50 identified duplicates, the manual analysis proved to be inline with the GeoNames dataset results. Out of the 50 duplicates, 25 had a 100% identical designation which means that they were most likely the same place imported twice from different sources. Furthermore, 20 of the selected duplicates had very similar spellings to the point that they could safely be considered to be different designations to the same place entity. These included examples such as *Nanacatepec* and *Nanacatepeque*, where the *c* at end is replaced by the similar sounding *que*, or in the use of grammatical accents in pairs like *Chimalhuacan* and *Chimalhuacán*. Lastly, the remaining 5 duplicates, while indeed similar, could not be considered duplicates with complete certainty and therefore were considered as unknown for this evaluation.

In the end, this test 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 in the random 50 duplicates selected from the gazetteer database. As such, it is also reasonable to assume that both scenarios share the lower specificity value meaning that there is a possibility that some duplicate entries were not correctly marked as such. This is an area for improvement although the choice to make sure that distinct places are not wrongly considered as duplicates should still be prioritized even if this means that some real duplicates are not flagged.

In short, this section demonstrated that the duplicate detection tool performs to a satisfactory degree and is especially accurate when dealing with minor alterations in the strings composing both place names. However, the low specificity values raises some questions regarding possible duplicates being missed due to the more conservative approach of the tool on marking them as such.

5.3.2 Polygon Data Generation Results

Like previously stated, this section presents scenarios which highlight the strengths and limitations of this tool in order to give a comprehensive overview of its capabilities during the data integration process and in possible future applications. In particular, the tool was used to recreate regions using two scenarios: (i) regions for which polygons were already stored in the database or (ii) regions for which their polygons

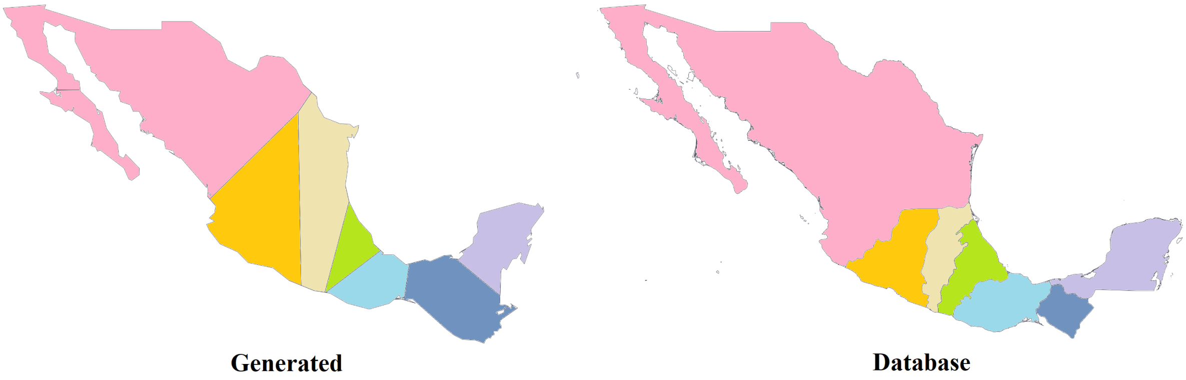


Figure 5.8: Side by side comparison of 16th century Mexican Diócesis.

were available from external sources. The similarity between the generated and the real geometries determined the overall accuracy and quality of the tool.

In the first scenario, the polygon generation tool attempted to recreate geometries for the Diócesis regions. Reference points of each Diócesis were retrieved from the database such as their respective boundary boxes and possibly some related points contained within them, which were used as input. Figure 5.8 presents a side by side comparison between the resulting generated dioceses and their original representation.

At a glance, the polygon generation tool managed to recreate the overall distribution of the original Diócesis regions. However, there is a major difference in the northern pink polygon which is considerably smaller in the generated version. It is likely that the tool overestimated the weights given to the orange and light yellow polygons which in turn caused both of them to cover a portion of territory which should belong to the pink polygon. The same situation is also probably at play with the dark-blue region entering into the light-purple and light-blue regions in the southern area although to a much smaller degree than the previous example. Despite these small inconsistencies the polygon generation tool managed to recreate the Diócesis to an acceptable degree compared to their original representation, providing a recreation which keeps the main divisions of the original.

The second scenario followed the same method of recreating a set of regions present in the database but this time with the geographical divisions of Mexico (i.e., Southern Mexico, Northern Mexico, Central Mexico, etc). The visual comparison between both sets of regions is present in Figure 5.9.

The generated outcome managed to maintain the overall shape and distribution of the original polygons. Compared to the first example the resulting polygons maintained their overall size and territory to a greater degree. However, the main shortcoming of the generated data in this example is related to the shape of each of the resulting polygons which changed compared to the original representation. In particular, the green region in the center suffered the most mainly due to its more inconsistent shape. It can then be concluded that the polygon generation tool has some limitations in accurately recreating

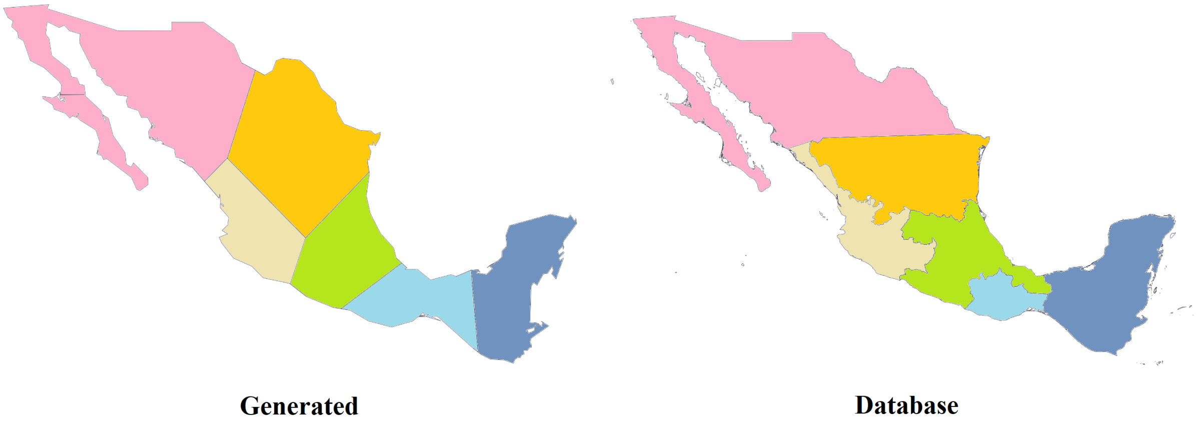


Figure 5.9: Side by side comparison of Mexican Geographical Regions

regions corresponding to irregular shaped polygons. Once again the generated polygons, while not a completely faithful recreation of the real regions, still managed to capture the general regional division to an acceptable level.

The majority of the other results of the tool are similar to the previous shown examples. The trend continues of correctly dividing the covered area into the correct number of regions with some issues related to the recreation of irregular shapes and the occasional disparity in weights causing an enlargement of some generated polygons.

Another more aggressive example of this weight analysis causing the wrong creation of larger polygons is present in the third example of Figure 5.10. The Figure shows an attempt to recreate the 16th century Gobiernos (i.e., the administrative divisions of colonial Spain) with quite large differences in size between each of its polygons, particularly in the western side of the territory. The obvious abnormality in the generated side is the black polygon which has a much higher area than it was supposed to. Further

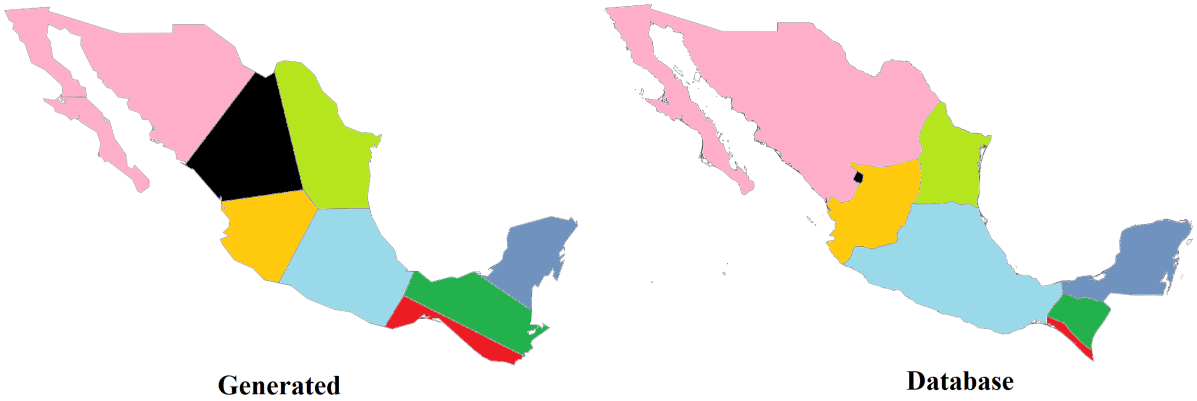


Figure 5.10: Side by side comparison of 16th century Mexican Gobiernos

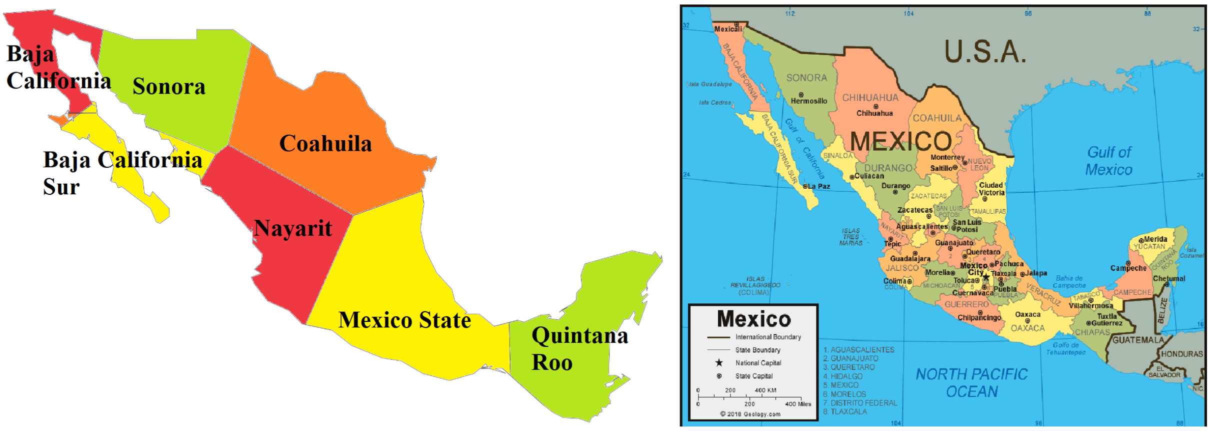


Figure 5.11: Side by side comparison of generated states with map of the Mexican states from Geology[3]

South, the red polygon largely remained inside its borders although it still occupied some of the light blue's area.

The attribution of an unit of area to a specific polygon is based around the combination of of the shortest distance to the main reference points with a weighed value based on the population of that location (i.e., higher population means higher priority for a reference point). It is possible to improve these results by either adding more relevant variables to the weight calculation or adding an entire new set of weights based on other parameters. At the time of writing this report this state of affairs was the best achievable performance regarding this issue.

Additionally, there is one more limitation worth mentioning for this section related to the lack of a complete list of regions when creating polygon divisions. Figure 5.11 presents a situation where this limitation can be observed. Unlike the previous examples, this scenario makes use of generated spatial footprints based around information from places of the type states which lacked a specific spatial footprint. However, the database only contained 6 states in this list which in turn caused the algorithm to be used with a clear lack of places to draw references from. Of course, as shown in the Figure, the generated polygons are far from the real state division. Some states such as Baja California Norte, Baja California Sul and Sonora did get a very accurate representation in the generated output but the same cannot be said for the remaining states, in particular, the central and southern states of Nayarit, México and Quintana Roo.

These results likely result from the polygon generation tool attributing area from other missing states to the nearest reference points in the list. Using an incomplete list of polygons can be concluded to have a negative influence in the outcome of the polygon generation since the algorithm tries to attribute area from missing real regions to the limited number of places used as input. This limitation of the algorithm requires entire new additions or reworks to its base logic in order to improve the outcome.

[3] https://geology.com/world/mexico-satellite-image.shtml

Overall, the polygon generation tool proved to be an overall accurate tool in the data integration process, that is capable of generating and attribute polygon regions to a reasonably satisfactory degree. With a complete list of places of the same type, each with a comprehensive list of reference points to be used as input, the outcome results in polygon regions is similar to the real regions corresponding to that same list of places, providing a reliable alternative in for places without a defined spatial footprint.

6

Conclusions and Future Work

Contents

6.1 Conclusions and Contributions

The main motivation behind this thesis consisted on the creation of a digital gazetteer framework capable of handling the many challenges of data management and integration present across multiple gazetteer projects. To achieve this, a set of initial objectives were established at the start of this document upon which the development of the gazetteer framework was based around. Chapter 3 explored a number of related projects whose contributions were implemented into the solution of the gazetteer framework such as the database schema of the ADL [7] or the data export format of the WHG [8] project. Through this initial analysis it was possible to create a working historical gazetteer framework which successfully implements a database for place data management and a web application that allows interaction in the form of search queries or export data in a multitude of formats.

Additionally, a complete process capable of importing and integrating geographical data from multiple sources into the gazetteer framework was developed. The results presented in Chapter 5 showcased a gazetteer database containing a large amount of place data which effectively represented the scope of the sources used to populate it. This includes handling temporal information of the places stored as well as establishing a large network of inter relations which better contextualizes all the places in the database. A more in depth analysis of the data present in the gazetteer showed that it was inline with the context surrounding its scope and other external sources. This in turn gives credibility to the process of importing and integrating data into the gazetteer database.

An analysis of the main tools used to perform the data integration process such as the duplicate detection and polygon generation tools also helped in evaluating the process of creating the gazetteer database. The duplicate detection tool showed promising results in detecting duplicates on a large sample of duplicate examples with these results also being notable on the gazetteer itself from randomly selected marked duplicates. In particular, it presented a tendency to being on the safe side when labelling duplicates that made the marked duplicates reliable at the expense of possible letting a number of real duplicates slip by. Meanwhile, the polygon generation tool managed to infer spatial footprints for a large number of vague places lacking such information. The results of this tool showcased its ability to accurately generate polygons which closely resembled the distribution of their real regions. However, some clear limitations for this tool were also identified such as handling neighbouring regions of vastly different sizes (the result tented towards balancing the size of both leading to incorrect polygons), accurately recreate polygons of irregular shapes and difficulties dealing with missing information (missing a couple of states when trying to generate that same type). Despite some shortcomings both of these tools proved valuable in the data integration process and consequentially improved the overall quality of the data that could then be exported through the web application interface.

In conclusion, the gazetteer created during the development of this project is in itself the main contribution of this MSc thesis. The combination of all its components and functionalities fulfils its intended role as a framework capable of importing and integrating gazetteer data from external sources in an effective way, and providing that data in a number of universal formats and representations.

6.2 Future Work and Improvements

The final section of this thesis will focus on listing possible improvements and future upgrades to the final gazetteer. A large number of possible changes can be proposed to the various different components.

Firstly, new types of export data formats can be added to the web application interface in order to expand the amount of ways upon which the gazetteer can provided its data to users and other projects. Other general improvements can also be made to the web interface such as a more complex search system with filters and multiple selection options which are now limited to the search of a place by its primary name.

The gazetteer's quality itself can still be improved by importing additional geographical data from other sources and through the inclusion of more annotations describing each place. A possible method to achieve this could be the use of crowd sourced data through its web application interface on the same veins of the GB1900 [23] project.

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 the Santos [1] neural network based string matching mechanism that could further improve the results of the integration process. Meanwhile, the polygon generation tool could also be improved by tackling the limitations described in the results section (e.g., handling adjacent different sized regions). A possible solutions would be to make use of contributions from related works in the area such as the web search based vague place delineation proposed by Jones [29] or one-class supervised learning through photo images of Cunha's [30] work.

Bibliography

[1] R. Santos, P. Murrieta-Flores, P. Calado, and B. Martins, “Toponym matching through deep neural networks,” *International Journal of Geographical Information Science*, vol. 32, no. 2, pp. 324–348, 2018.

[2] S. Gao, L. Li, W. Li, K. Janowicz, and Y. Zhang, “Constructing gazetteers from volunteered Big Geo-Data based on Hadoop,” *Computers, Environment and Urban Systems*, vol. 61, pp. 172 – 186, 2017.

[3] R. S. Purves, S. Winter, and W. Kuhn, “Places in information science,” *Journal of the Association for Information Science and Technology*, vol. 70, no. 11, p. 1173–1182, Mar 2019.

[4] L. Berman, R. Mostern, and H. Southall, “On historical gazetteers,” *International Journal of Humanities and Arts Computing*, vol. 5, pp. 127–145, Oct 2011.

[5] E. Acheson, S. De Sabbata, and R. Purves, “A quantitative analysis of global gazetteers: Patterns of coverage for common feature types,” *Computers Environment and Urban Systems*, vol. 64, p. 309–320, 07 2017.

[6] L. Berman, *Placing Names: Enriching and Integrating Gazetteers.* Indiana University Press, Aug 2016.

[7] L. Hill, J. Frew, and Q. Zheng, “Geographic names,” *D-Lib Magazine*, vol. 5, no. 1, 1999.

[8] P. Manning, “World-historical gazetteer,” *Journal of World-Historical Information*, vol. 2, Aug 2015.

[9] R. Lake, D. Burggaf, M. Trninic, and L. Rae, *Geography mark-up language: foundation for the geo-web*. Wiley, 2004.

[10] R. Santos, P. Murrieta-Flores, and B. Martins, “Learning to combine multiple string similarity metrics for effective toponym matching,” *International Journal of Digital Earth*, vol. 11, no. 9, pp. 913–938, 2018.

[11] V. I. Levenshtein, “Binary codes capable of correcting deletions, insertions and reversals.” *Soviet Physics Doklady*, vol. 10, no. 8, pp. 707–710, feb 1966, doklady Akademii Nauk SSSR, V163 No4 845-848 1965.

[12] F. J. Damerau, “A technique for computer detection and correction of spelling errors,” *Communications of the ACM*, vol. 7, no. 3, pp. 171–176, Mar. 1964.

[13] M. Jaro, “Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida,” *Journal of the American Statistical Association*, vol. 84, no. 406, 1989.

[14] W. Winkler, “String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage,” *American Statistical Association*, vol. 84, no. 406, 1990.

[15] P. Jaccard, “The distribution of the flora in the alpine zone.1,” *New Phytologist*, vol. 11, no. 2, pp. 37–50, 1912.

[16] A. E. Monge and C. P. Elkan, “The field matching problem: Algorithms and applications,” in *Proceedings of the International Conference on Knowledge Discovery and Data Mining*, 1996.

[17] Edelsbrunner, D. G. Kirkpatrick, and R. Seidel, “On the shape of a set of points in the plane,” *IEEE Transactions on Information Theory*, vol. 9, 1983.

[18] B. N. Delaunay, “Sur la sphère vide.” *Bulletin de l’Académie des Sciences de l’URSS*, vol. 1934, no. 6, pp. 793–800, 1934.

[19] G. Voronoi, “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)*, vol. 1908, no. 134, p. 198–287, Jan 1908.

[20] P. Harpring, “Development of the getty vocabularies: AAT, TGN, ULAN, and CONA,” *Art Documentation: Journal of the Art Libraries Society of North America*, vol. 29, no. 1, pp. 67–72, 2010.

[21] R. Simon, L. Isaksen, E. Barker, and P. d. S. Cañamares, “Peripleo: a tool for exploring heterogeneous data through the dimensions of space and time,” *Code4Lib*, no. 31, Jan 2016.

[22] L. Isaksen, R. Simon, E. T. Barker, and P. de Soto Cañamares, “Pelagios and the emerging graph of ancient world data,” in *Proceedings of the ACM Conference on Web Science*, 2014.

[23] H. Southall, P. Aucott, C. Fleet, T. Pert, and M. Stoner, “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*, vol. 13, no. 1, p. 7–28, Jan 2017.

[24] J. T. Hastings, “Automated conflation of digital gazetteer data,” *International Journal of Geographical Information Science*, vol. 22, no. 10, pp. 1109–1127, 2008.

[25] P. Frontiera, R. Larson, and J. Radke, “A comparison of geometric approaches to assessing spatial similarity for GIR,” *International Journal of Geographical Information Science*, vol. 22, no. 3, pp. 337–360, 2008.

[26] F. Rosenblatt, *The Perceptron, a Perceiving and Recognizing Automaton Project Para*. Cornell Aeronautical Laboratory, 1957.

[27] D. W. Goldberg, J. P. Wilson, and C. A. Knoblock, “Extracting geographic features from the internet to automatically build detailed regional gazetteers,” *International Journal of Geographical Information Science*, vol. 23, no. 1, pp. 93–128, 2009.

[28] L. See, P. Mooney, G. Foody, L. Bastin, A. Comber, J. Estima, S. Fritz, N. Kerle, B. Jiang, M. Laakso, and et al., “Crowdsourcing, citizen science or volunteered geographic information? the current state of crowdsourced geographic information,” *ISPRS International Journal of Geo-Information*, vol. 5, no. 5, p. 55, Apr 2016.

[29] C. B. Jones, R. S. Purves, P. D. Clough, and H. Joho, “Modelling vague places with knowledge from the Web,” *International Journal of Geographical Information Science*, vol. 22, no. 10, pp. 1045–1065, 2008.

[30] E. Cunha and B. Martins, “Using one-class classifiers and multiple kernel learning for defining imprecise geographic regions,” *International Journal of Geographical Information Science*, vol. 28, no. 11, pp. 2220–2241, 2014.

[31] B. Plewe, “The nature of uncertainty in historical geographic information,” *Transactions in GIS*, vol. 6, no. 4, p. 431–456, 2002.

[32] C. Freksa, “Temporal reasoning based on semi-intervals,” *Artificial Intelligence*, vol. 54, no. 1, pp. 199 – 227, 1992.

[33] T. Kauppinen, G. Mantegari, P. Paakkarinen, H. Kuittinen, E. Hyvönen, and S. Bandini, “Determining relevance of imprecise temporal intervals for cultural heritage information retrieval,” *International Journal of Human-Computer Studies*, vol. 68, no. 9, pp. 549 – 560, 2010.

[34] S. Brandon and Plewe, “Representing Datum-level Uncertainty in Historical GIS,” *Cartography and Geographic Information Science*, vol. 30, no. 4, pp. 319–334, 2003.

[35] M. Wick and B. Vatant, “The geonames geographical database,” *Available from World Wide Web: http://geonames.org*, 2012, accessed: Dec 2020.