Digital gazetteers: review and prospects for place name knowledge bases

KALANA WIJEGUNARATHNA and KRISTIN STOCK, School of Mathematical and Computational Sciences, Massey University, New Zealand
CHRISTOPHER B. JONES, School of Computer Science and Informatics, Cardiff University, UK

Gazetteers typically store data on place names, place types and the associated coordinates. They play an essential role in disambiguating place names in online geographical information retrieval systems for navigation and mapping, detecting and disambiguating place names in text, and providing coordinates. Currently there are many gazetteers in use derived from many sources, with no commonly accepted standard for encoding the data. Most gazetteers are also very limited in the extent to which they represent the multiple facets of the named places yet they have potential to assist user search for locations with specific physical, commercial, social or cultural characteristics. With a focus on understanding digital gazetteer technologies and advancing their future effectiveness for information retrieval, we provide a review of data sources, components, software and data management technologies, data quality and volunteered data, and methods for matching sources that refer to the same real-world places. We highlight the need for future work on richer representation of named places, the temporal evolution of place identity and location, and the development of more effective methods for data integration.

CCS Concepts: • **General and reference** → **Surveys and overviews**; • **Information systems** → *Geographic information systems*; *Information retrieval*.

Additional Key Words and Phrases: Gazetteers, Toponyms, Place Names, Geographic Information Retrieval, Information Retrieval, Geographic Databases, Geographic Information Systems

ACM Reference Format:
Kalana Wijegunarathna, Kristin Stock, and Christopher B. Jones. 2025. Digital gazetteers: review and prospects for place name knowledge bases. 1, 1 (August 2025), 39 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Gazetteers have a long history of use to represent information about places. Digital gazetteers usually take the form of databases or knowledgebases that store information about place names (also referred to as toponyms) and support access to many geospatial applications. They may also be referred to as geographical gazetteers to distinguish them from the less common use of the term gazetteer in natural language processing, to refer to a dictionary or a list of names that includes places, but also organizations, people, and domain-specific terms [111, 131]. The earliest geographical gazetteers[1] listed place names and described places with regard to their physical and

[1]The word gazetteer dates back at least to the seventeenth century, to describe either a journalist or a newspaper, while its use to refer to a geographic index of places became more common in the nineteenth century (Oxford English Dictionary https://www.oed.com/).

Authors' Contact Information: Kalana Wijegunarathna, k.wijegunarathna@massey.ac.nz; Kristin Stock, k.stock@massey.ac.nz, School of Mathematical and Computational Sciences, Massey University, Auckland, New Zealand; Christopher B. Jones, jonescb2@cardiff.ac.uk, School of Computer Science and Informatics, Cardiff University, Cardiff, UK.

ACM XXXX-XXXX/2025/8-ART
https://doi.org/10.1145/nnnnnnn.nnnnnnn

social characteristics. Ptolemy's *Geographia* appears to be one of the earliest recorded gazetteers (although it is a revision of a now lost atlas believed to have predated it), and was primarily used for navigation [15]. Digital gazetteers tend to be considerably less comprehensive than historical gazetteers, often lacking extensive place descriptions (see the Gazetteer for Scotland[2] for a rare exception). However, they almost always include geographical coordinates (e.g., latitude and longitude) along with the type of place (e.g. river, town), referred to in this paper as geographic feature type. Additional elements can include alternative place names, demographic information, and details of the place name etymology. With this shift, gazetteers are now often used to support place name-based search, such as in web mapping and navigation, or to support geospatial referencing (geoparsing and georeferencing) of text documents [68].

Gazetteers have become a cross-cutting research topic encompassing Geographic Information Science (GIS), Information Retrieval (IR), computer science, and history. They play a key role in systems and research that exploit or depend upon the use of place names to access or manage information. For example, in scenarios involving web or social media searches, web mapping services, or digital library queries aimed at locating geographic information, a gazetteer could be employed to recognise and disambiguate (find the correct instance of) a place name in the query, and use the resulting coordinates to pan and zoom to the relevant location on a map [25, 27, 47, 122, 128, 137, 138, 142, 154]. Routing services for emergency response, public transport or navigation could also employ gazetteers to recognise user-specified place names. The utilization of social media content for disaster response may require a gazetteer to locate named places in the disaster zone. The processes of toponym recognition and toponym resolution (disambiguation and geocoding) are relevant both to understanding a spatial query and to analysing text to find geographical references, referred to as georeferencing. The presence of a candidate name in a gazetteer can be part of the evidence for identification as a toponym, while the disambiguation and geocoding process can use gazetteers to generate their coordinates, for example [88]. Current named entity recognition (NER) methods, of which toponym recognition is a sub-task, may dispense with a gazetteer, relying instead on learning from expansive corpora. Although language modelling methods have been developed to generate coordinates for toponyms or text documents, a gazetteer remains advantageous for generating precisely disambiguated coordinates [44, 105].

In addition to the above applications, gazetteers have been developed for applications in the humanities, for example to record names and locations of historical and archaeological artefacts and places of interest. An example is The World Historical Gazetteer (WHG) which can be searched temporally and is contributed to by archaeologists and historical researchers [61]. Temporal search of gazetteers can support study of the evolution of names [26, 45, 160]. Other history-related applications of gazetteers include [20, 71, 89, 164]. Gazetteers also act as a source for ethnophysiographic studies of how languages and cultures refer to natural landscape terms [97].

Transformations in computer science, information retrieval and technology have led to digital gazetteers evolving in some cases from being simple name lookups to becoming aspects of complex geospatial knowledgebases such as Google Maps[3] and OpenStreetMap (OSM) [10] that combine typical gazetteer functionality with storage of detailed digital map data. Many of the developments have led however to experimental gazetteers produced by researchers whose work has not been properly integrated or maintained. Notably, most widely used gazetteers have no common content standard or feature type schema - and the call [80] for an agreed feature type ontology has not been met. The result is a proliferation of gazetteers which, while potentially regionally comprehensive, have limited or no alignment with respect to whether their respective records refer to the same

[2]https://www.scottish-places.info/
[3]https://www.google.com/maps

place, or with regard to access methods. Whether it is possible or even desirable to follow a single standard for gazetteers is also a question that has been raised, due to the wide variety of both uses and users for gazetteers [67, 80]. Understanding the diversity of approaches to gazetteer design and potential for greater integration of gazetteer content is one of the main motivations behind our review of gazetteers, gazetteer technologies and standards.

While digital gazetteers have a wide range of applications, there is often little information about the places referred to by toponyms. This is in contrast to traditional pre-digital gazetteers that, as indicated above, often recorded a much broader range of information about a place including for example its historical origins, events that have taken place there, its architecture and types of services that it affords. The latter could include religious institutions, commercial companies and industrial factories. Numerous academic publications have considered the multiple dimensions or facets of the concept of place, including how they can be perceived by different people and hence, by implication, why there could be multiple reasons for particular places to be of interest [63, 123, 143]. There is a motivation therefore to maintain such facets in a digital gazetteer.

When people query web mapping systems they frequently combine a place name with a service or activity of interest, perhaps relating to food and drink, leisure activities, culture, sport, or some commercial enterprise. Given that the main general purpose digital gazetteers provide either no or limited support for the service and affordance aspects of such queries, the search engines can be expected to depend upon curated data sources in the form of point of interest (POI) datasets, or on web pages that associate a place name with a service or activity. A major source of more substantial data about places is provided by the place-specific pages in Wikipedia and their semantic web presentations on DBpedia, but there is great variation in the content and geographic coverage of these pages. In conventional digital gazetteers the place type data item enables some description of the named place, but often this is recorded as only a single feature type. It may be argued that gazetteers could provide improved support for geospatial queries, and serve as more general purpose knowledge bases of named places, if they recorded much richer information about the place to which the name refers. The place could then be regarded as a multi-faceted information object that is the basis of a gazetteer record [123].

In this paper, we review publications addressing the topic of gazetteers published since the last decade of the 0^{th} century. We draw comparisons between papers through various aspects: reviewing the current standing of the standards for gazetteers; assessing the numerous methods for identifying, deduplicating and disambiguating place names; integrating place name sources; identifying gaps and potential for improvement. We also review the data items commonly found in gazetteers, and we highlight the distinction, indicated above, between gazetteers as quite sparse representations of place name knowledge and the potential for gazetteers to reinstate more widely their standing as rich representations of knowledge about the named places. Our intention is to provide insights that will help future gazetteer developers make decisions on these areas. We also propose areas for further research and areas that may benefit from the integration of the latest developments in computer science, particularly those of deep learning and AI.

We will address these research questions in this paper:

(1) What are the core components of a gazetteer and how have they evolved over the years?
(2) What place name data sources have been used to populate gazetteers?
(3) What methods have been most effective in the compilation and integration of gazetteers?
(4) What implications has the growth of the web and Volunteered Geographic Information (VGI) had on the building of gazetteers?
(5) What technologies have been used for the implementation of gazetteers?
(6) What are the limitations of digital gazetteers and how can they be addressed?

The remainder of the paper is structured as follows: In the next section, we summarise previous relevant reviews in the area. In Section 3, we present the methodology for the literature review. In Section 4, we describe the evolution and classification of gazetteers, followed by a discussion of the contents of a gazetteer in Section 5. In Section 6, we discuss the sources used to build gazetteers and the process of integration of these sources. Section 7 discusses VGI and gazetteers on the web, and Section 8 reviews gazetteer technologies. In Section 9, we discuss the improvements that can be made to gazetteers to suit modern applications and avenues for future research, while Section 10 sumarises conclusions.

2 Related Studies

Upon conducting an extensive search, we were unable to identify publications primarily focused on reviewing the existing literature on digital gazetteers. Nonetheless, several papers that address the broad topic of the construction and evolution of gazetteers were found, some of which consist of a substantial review component. These findings are summarized here.

[68] discusses gazetteers in light of the Alexandria Digital Library (ADL) gazetteer. A gazetteer is defined as a geospatial dictionary of geographic names with the core components of a name, a location, and a type. This definition is one of the most frequently cited of a modern gazetteer. The paper discusses the state of gazetteers at that point in time, elaborating on the three main components, with particular regard to the ADL, and discussing underlying principles relevant to place names. The paper provides an informative introduction to gazetteers in the late 20^{th} century.

In proposing future directions and challenges for gazetteers, [80] identify shortcomings of existing gazetteers and upcoming trends, particularly the incorporation of VGI, discussing challenges and solutions. A more recent article on gazetteers [58] discusses the prominence of VGI as a key component of *neogeography* [144], the phenomenon of people, particularly outside the discipline of geography, creating and interacting with digital geographic information and maps. Even though their interest is not in reviewing gazetteers, they do provide an introduction to gazetteers.

[117] aim to guide new gazetteer builders in making design decisions based on the authors' experience. Though they discuss the merits of particular approaches to building and publishing gazetteers (such as Linked Open Data), they do not present a current standing of gazetteers or study their components in depth.

None of these papers has comprehensively discussed the evolution of gazetteers, their components, and the application of latest technologies in constructing a gazetteer, as we aim to here.

3 Methodology

In order to answer our research questions, we designed two search queries to run on three main source databases: Scopus, Web of Science and EBSCO Discover. The search queries are as follows:

(1) (TITLE(gazetteer*) AND TITLE-ABSTRACT-KEYWORDS(digital OR geograph*)) OR (TITLE(place names) AND TITLE-ABSTRACT-KEYWORDS(gazetteer*))
(2) TITLE (place name database OR place name ontology OR place name knowledge base OR place name knowledgebase OR toponym database OR toponym ontology OR toponym knowledgebase OR toponym knowledge base)

The first query was crafted to encompass as many papers as possible while limiting the focus to geographical gazetteers. However, this approach may have inadvertently excluded papers where gazetteers were employed but referred to by alternative terminology, or were included within wider knowledge bases, ontologies, or databases, and hence the second query was also included. Both queries were limited to journal papers and peer reviewed conference papers published from the year 2000 to 2022. We then enhanced the searches with papers identified through examination

of the reference lists of relevant papers. This helped us to capture papers that were not directly retrieved from our search queries including some before the year 2000. Among papers resulting from our initial search queries, we came across some that were not directly relevant to the topic or answered any of the RQs. Therefore, we excluded papers that met the following exclusion criteria:

(1) Papers that did not use gazetteers.
(2) Papers discussing older gazetteers which were not exclusively geographic but more administrative catalogs.
(3) Papers that used the term "gazetteer" to refer to dictionaries or name records (which may or may not include place names).
(4) Papers that publish simple lists of place names

A breakdown of the number of papers extracted is shown in Fig. 7 (Appendix).

A few other references relevant to our research questions are included, which were not in the search results or cited by them, but which the authors were aware of or were found through searches using Google Scholar. Google Scholar was not included as a source in the first search due to the broad range of materials that it includes that may not be peer-reviewed.

4 The Evolution and Classification of Gazetteers

Many historical gazetteers were created by Kings, Emperors or rulers of kingdoms and contained more information than is commonly found in contemporary gazetteers. Though often lacking coordinate pairs or other forms of geographic footprints, these gazetteers commonly include information such as population, descriptions of the boundaries of the place, landmarks that made the identification of the place easier, historically well-known people or incidents, and administrative information such as records of tax collection and commodities [107]. For example, the Domesday Book recorded an extensive survey of Britain in 1086, listing land area; number of people (including slaves); livestock; income and taxation (see Fig. 1). From the late eighteenth century, similar forms of gazetteers became increasingly common, sometimes also including local information about facilities such as train stations and schools (see for example Wilson's Gazetteer of Scotland [156]).

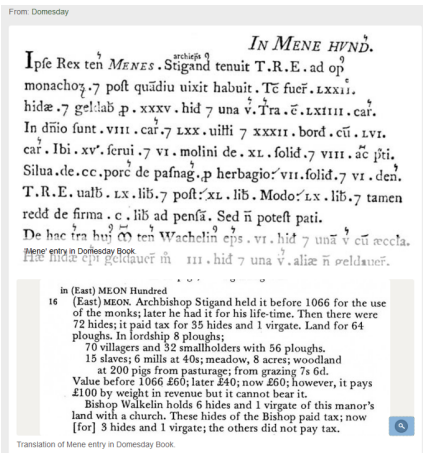

Fig. 1. Example of content from Domesday Book (Source: https://www.eastmeonhistory.org.uk/content/catalogue_item/domesday/mene-domesday-book. Translated by Gordon Timmins).

In the mid twentieth century, gazetteers adopted a more quantitative approach to spatial information, focusing on collection and storage of coordinates or other spatial footprints, feature types

and administrative divisions that allowed the location of a place to be specified accurately [79]. Historically, gazetteers were primarily curated by national mapping authorities and detailed the place names within a specific country. The first initiative to compile a global gazetteer by aggregating individual national gazetteers was undertaken by the Economic and Social Council of the United Nations (ECOSOC) in the 1950s, although the concept was initially proposed at the Fifth International Geographical Congress in 1891. In the 1960s and 70s, the United Nations Group of Experts on Geographical Names (UNGEGN) passed several resolutions with various recommendations about definitions, standards and resolution of national differences. However, the envisaged world gazetteer did not come to fruition due to differences among countries [133]. Nevertheless, initiatives taken by individual countries became imperative to the development of various gazetteer-like services in the last few decades of the twentieth century [79].

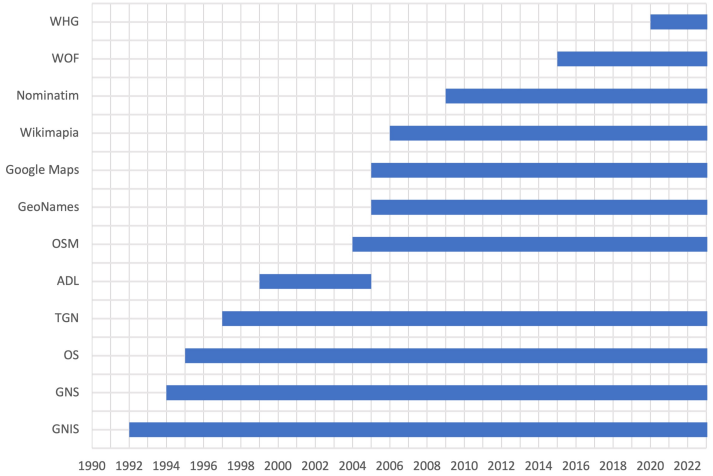

Fig. 2. Timeline of Commonly Referenced Gazetteers. (see text for acronyms)

Fig. 2 shows a timeline for some well-known, commonly used gazetteers. Not all were returned by our search queries (as they are internet applications), but they are examples of popular gazetteers (or services that act as gazetteers even though named otherwise). The list is not exhaustive. All of them exist or have existed as commercial or publicly available place name information sources.

The Geographic Names Information System (GNIS)[4] is a place name database managed by the U.S. Board on Geographic Names (BGN)[5], the authoritative body responsible for place names in the U.S. The GeoNET Names Server (GNS)[6] was created by the U.S. National Geospatial-Intelligence Agency (formerly the National Imagery and Mapping Agency) and was originally responsible for collecting the names of places outside the U.S. These gazetteers are listed as the sources for the Alexandria Digital Library Gazetteer (ADL). Even though the ADL gazetteer was only in operation for a short period of time, the developers attempted to build a common and comprehensive gazetteer standard known as the Gazetteer Content Standard (GCS), along with the ADL Feature Type Thesaurus (FTT) type scheme for places (see Section 5.2). The GCS has not been adopted as a de facto standard, but it incorporates a wide range of data items including alternative names,

[4]https://www.usgs.gov/tools/geographic-names-information-system-gnis
[5]https://www.usgs.gov/us-board-on-geographic-names
[6]https://geonames.nga.mil/geonames/GeographicNamesSearch/

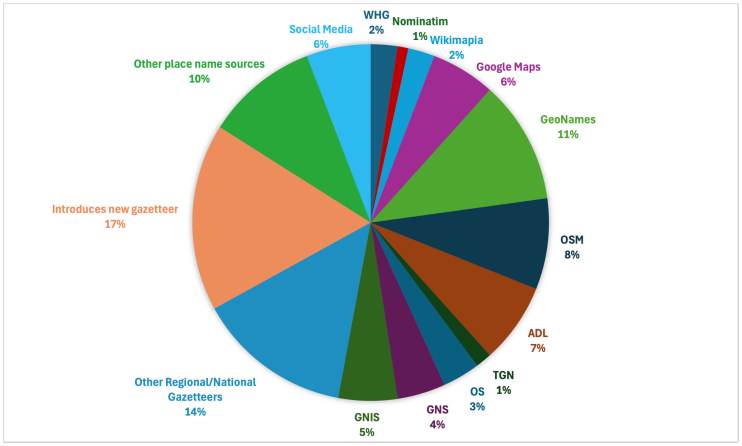

Fig. 3. Usage of the prominent gazetteers in retrieved articles. Figure also presents articles that introduce new gazetteers and source non-gazetteer sites as place name information sources.

temporality, multiple geometries, pronunciation and etymology of names, certainty of attributes and provenance information. Due to the scarcity of data, a significant number of records in the ADL gazetteer lacked many of the proposed GCS attributes. We discuss the GCS, the FTT and the ADL gazetteer in light of specific topics throughout the paper.

Britain's Ordnance Survey (OS)[7] completed the digitization of all its maps in 1995 along with an associated gazetteer. It was accompanied by the development of many digital gazetteers listing property addresses managed by local government agencies. These property and street gazetteers were subsequently managed jointly by a centralized agency GeoPlace[8]. The British Ordnance Survey is just one of many national mapping agencies that have produced gazetteers or place name databases in association with their map products. The Getty Thesaurus of Geographic Names (TGN) [9] was compiled and maintained by the Getty Research Institute. It does not call itself a gazetteer, but it is a place name knowledge base with many geographic footprints [68, 72].

The early twenty-first century saw a prevalence of open gazetteers that use publicly volunteered (crowdsourced) information, including OpenStreetMap (OSM)[10], GeoNames[11] and Wikimapia[12]. OpenStreetMap is associated with the Nominatim[13] gazetteer tool to access the place names in the OSM database. Google Maps is a large commercial undertaking launched in 2005 and Google Places is a service that provides place name information. These products that had their inception in the early 2000s now have significant coverage of the earth with millions of data points. Who's on First (WOF)[14] is a recent project to build a global gazetteer. The World Historical Gazetteer (WHG)[15] is a gazetteer listing historical places to aid historians and other humanities researchers to map historical places, artefacts, and events.

[7]https://www.ordnancesurvey.co.uk/
[8]https://www.geoplace.co.uk
[9]https://www.getty.edu/research/tools/vocabularies/tgn/index.html
[10]https://www.openstreetmap.org/
[11]https://www.geonames.org/
[12]https://wikimapia.org/
[13]https://nominatim.org/
[14]https://whosonfirst.org/
[15]https://whgazetteer.org/

We present the frequency of usage of these prominent gazetteers in the articles retrieved in Fig. 3. In addition to the application of existing gazetteers, we also report here the articles that introduce new gazetteers, and articles that harvest various other sites, notably social media, as place name sources (e.g. Flickr, Panoramio, Twitter). Other non-gazetteer resources that have been treated or harvested as place knowledgebases (e.g. Wikipedia, bio-diversity collection records) are recorded under the "Other place name sources" category. While some prominent and more frequently used national gazetteers, such as the British OS and GNIS, have been included separately, many other national or regional gazetteers were identified (e.g. SwissTopo, China Historical GIS (CHGIS), Norwegian Central Place Name Register (SSR), Chinese National Geomatics Center's gazetteer). This category of "Other Regional/National Gazetteers" also includes ancient gazetteers which are almost always regional. "Google Maps" includes both Google Maps and the Google Places API. In the context of this figure, the use of a gazetteer could be to create a new gazetteer, the use of places from a gazetteer for the creation of datasets to test (and or train) new models or methods, or an analysis of a gazetteer (e.g. articles discussing the coverage or data quality of specific gazetteers).

Free and crowdsourced gazetteers like GeoNames and OSM are the most commonly used established gazetteers. This observation is quite intuitive given their easy accessibility, but also indicates the level of confidence placed in crowdsourced data and the measures taken by these gazetteers to manage the quality of VGI. Another observation of interest is the absence of any papers (among those retrieved) that conducted research using Google Maps/Google Places API since 2018, when Google introduced a significant overhaul to its Maps Platform pricing model, making it no longer freely available for developers. Despite being discontinued, the ADL gazetteer remains a prominent knowledge source not only as a gazetteer itself (when it was available) but for the ADL GCS. This reflects its original intentions for building a common gazetteer standard and data model. The large number of gazetteers that are introduced in the papers, however, hints at the acute challenge of defining and gaining adoption for such a standard.

Over the years the feature types of places in gazetteers have been encoded with various knowledge representation schemes in the semantic spectrum. The semantic spectrum [112] ranks different semantic technologies according to their semantic strength and interoperability. Even though according to Obrst [112], the semantic spectrum ranks ontologies, it often also includes simpler and less semantically powerful Knowledge Organization Systems (KOS) like lists, dictionaries, and controlled vocabularies [112]. It is beyond the scope of this paper to discuss the semantic spectrum in detail, but it is vital to understand some of the knowledge organisation systems used in gazetteers. We will discuss feature typology in detail in Section 5 but introduce some KOS here since they play a vital part in the classification of gazetteers. Some gazetteers use these systems for defining place types while there are others that use them as their data structure to maintain all contained place names.

The simplest models are controlled vocabularies - simply a finite collection of items. A glossary is slightly more advanced using free text to describe the meanings of terms. Increasing semantic interoperability leads to thesauri and taxonomies. A thesaurus provides synonyms, broader/narrower terms and association relations for its terms. Taxonomies provide stronger hierarchical relations in addition to the features provided by a thesaurus. More semantically enriched are conceptual models and logical theories. A conceptual model is considered a weak ontology and a logical theory is a strong one [112], where an ontology is defined as an explicit specification of a conceptualization [62]. A conceptualization is made up of a set of objects, concepts, and relations between them about which knowledge is being expressed. A conceptual model has generalised relations (class level), properties, instances, and attributes, while a logical theory is enhanced further with axioms and rules. Logical theory enables machine semantic interpretation as it is represented in a

logical knowledge representation. Both these types of ontologies can have concept attributes, and different types of relations.

Gazetteers use different semantic models from various levels in the spectrum. Early digital gazetteers used controlled vocabularies and thesauri (GNIS, GNS, ADL, GeoNames). The beginning of the 21st century saw a trend toward use of web-based ontologies, discussed in the section on Gazetteer Technologies. In parallel with the use of ontologies was the use of semantically enriched typology exemplified by folksonomies. Folksonomies can be regarded as a type of taxonomy where the public tags online items. The folksonomy used by OSM enforces a loose hierarchy but users or the public who add a place to OSM are allowed to tag the places with their own place types - leading to a large number of tags whose similarities or differences, even within the same hierarchy are sometimes hardly discernible. We discuss the use of feature type models in detail in Sections 5.2 and 8.

5 Components of a Gazetteer

The three main components of a gazetteer identified earlier: place name, feature type and geographic footprint, have remained the backbone of modern gazetteers. Depending on the gazetteer's use cases, various other changes and additions have been proposed to make them more useful along with various extensions to the three main components.

5.1 Place Names

Place names are an essential component of a gazetteer as they provide the most common means by which people refer to locations. The use of place names however introduces challenges of ambiguity, duplication, multilingualism and the need to resolve and integrate local and vernacular place names. [85] discusses these challenges referring to each of the issues we present with examples in Table 1.

The ADL gazetteer content standard (GCS) provided support to address some of the above issues, particularly with regard to enabling encoding of alternative versions of the name of geographic feature. It also specified a range of attributes of each place name and its variants [69]. Listed below are some of them, but of these only time period was a required element (see also [67]).

- Official or authoritative source if any
- Etymology (derivation) of name
- Language code of the language the name was written in. (for example, 'en' for English; 'fr' for French; 'mi' for Māori)
- Pronunciation (text or audio file)
- Transliteration scheme
- Confidence in the name
- Abbreviations if any
- Time period. Three options – former, current and proposed
- Links/references to more information

Early digital gazetteers developed in the late 1900s compiled places extracted from authoritative sources. These sources included non-digital gazetteers, maps, and records from official toponymic authorities. In the first decade of this millennium the interest in web harvesting, and later the advent of VGI, meant gazetteers were no longer limited to authoritative sources, and hence could include non-official place names. This included vague place names and vernacular place names [77].

Table 1. Challenges in dealing with place names.

Challenge	Description	Example
Feature ambiguity	The same name could refer to two different features because they are of two different feature types.	New Zealand, the country, is different from New Zealand the group of islands referring to a physical entity.
Multilingualism	The same place being called different names in different languages.	The city of Auckland, New Zealand is also known as Tāmaki Makaurau, its original Māori name.
Vernacular names	Place names that are informal or colloquial, that can be alternative to formal, official or administrative names, or might be present in the absence of a formal name. Vernacular names can sometimes be vague with imprecise boundaries.	Brum to refer to Birmingham in England. Examples of vague vernacular names are The South of France, and Downtown Los Angeles.
Temporal changes	Names of places change with time with political, administrative or socio-ethic reasons.	Istanbul used to be called Byzantium and Constantinople.
Place name ambiguity	Two or more different places being called the same place name.	Lisbon is the capital of Portugal but there are also several towns in different states of the US called Lisbon. The state of Wisconsin, USA contains at least 4 inhabited palces called Springfield.
Entity ambiguity	A name could be shared by a geographic place as well as another prominent person, organization etc.	Liverpool could be referring to the city in England or the Liverpool football club.
Region specific names	The same feature (especially large natural features like mountains or rivers) may have different names depending on the different regions the feature crosses.	The river Danube is called “Donau” in Germany, “Dunaj” in Slovakia, “Duna” in Hungary, “Dunav” in Croatia, “Dunav” and “Дунав” in Bulgaria, “Dunărea” in Romania and in Moldova and “Dunaj” and Дунай” in the Ukraine.
Abbreviations	Abbreviations of names.	Los Angeles is abbreviated to L.A. or S.F. for San Francisco.
Name Sim--plifications	Similar to abbreviations but some places have simplifications for the name for reasons such as the original name being harder to pronounce or write.	Place with the longest place name – a place in New Zealand called “Taumat--awhakatangihangakoauauotamateatu--ripukakapikimaungahoronukupokai-whenuakitanatahu” is often simplified as “Te Taumata”.
Nicknames	Common nicknames are type of alternative names. Similar to a vernacular name, but is more likely to be used to embellish a description of a place, rather than be commonly used as an alternative to the official name.	The capital of France, Paris, is also known as the “City of light”.

Vernacular place names are names that are commonly used regardless of whether they are official or not. They can also include vague place names that refer to locations lacking distinct boundaries, or places whose extents and landmarks might be perceived differently by various individuals and across different contexts [39]. Their acquisition is vital for information retrieval tasks [147] in order to ensure such names are recognised when used in queries. GumTree [148], Craigslist [72], phonebooks sites [56], social media sites like Flickr and Panoramio [118] are examples of sources that have been harvested for vernacular names. Several methods employed in the extraction of vernacular place names from these websites can be identified throughout literature. One of the early methods was based on analysing content of scraped web documents using regular expressions to identify addresses and a list of spatial prepositions to detect place names, followed by a web search validation procedure [148]. They report about 20 place names obtained from GumTree, an advertising and community website, in and around Cardiff, UK, that were not found in the British Ordnance Survey OS50k gazetteer.

Vernacular names have also been extracted from web pages using targeted query methods. Thus queries could take forms such as "$\langle target\ region \rangle$", "$\langle concept \rangle\ \langle target\ region \rangle$" or "$\langle lexical\ pattern \rangle\ \langle target\ region \rangle$" [7, 77]. The target region is a name to be extracted, the concept can be a feature such as a hotel while the lexical pattern (also called a trigger phrase) could include spatial relations as "is [in | located in | situated in] the [center | north | south | east | west] of" where | means or. NER methods can then be used to detect place names in the retrieved snippets of text. An alternative, web-based approach to vernacular name detection is to mine the content of addresses on web pages, in which vernacular names of neighbourhoods can be found within the address structure, using regular expressions. The coordinates are then indicated by any accompanying postcode [24].

[72] use more sophisticated natural language processing techniques such as NER, i.e. the recognition of important nouns and proper nouns, on text scraped from Craigslist. They combine two off the shelf NER tools, spaCy[16] and Stanford NER[17] , with two more case sensitive and Twitter retrained Stanford variants [72]. They report their method can enhance gazetteers with new places (places that were not recorded in existing gazetteers) especially for features like local neighbourhoods, parks, schools and points of interest along with other alternative vernacular names for places already reported in gazetteers such as GeoNames, TGN and WOF.

[127] demonstrated the ability to rely fully on VGI for vernacular place names within a small area – the George Mason University. The gazetteer was constantly updated by students and staff with changes, repairs and closedowns within the university premises to help the visually impaired. A slightly different study was carried out to retrieve peoples' perception of places and the vernacular names they use in [146] where the authors created a web platform[18] for people to name places within a familiar geographical location. This was done variously with an interactive map to name all the place names they knew within an area that they specified, using a postal code, or asking for names used in place of official names. While this may be an effective method to collect vernacular names, its extensibility depended largely on the interest of the crowd to participate in it.

5.2 Feature Type

The feature type is defined in [68] as a type "selected from a type scheme of categories for places/features" [68, p.1], and examples include natural features such as a river or a mountain, artificial features like a building, post office or village, and the administrative class or jurisdiction. Gazetteers vary according to whether only a single type is recorded for a place name instance, or whether multiple

[16]https://spacy.io/

[17]https://nlp.stanford.edu/software/CRF-NER.shtml

[18]https://peoplesplacenames.com/map.php

types are sometimes listed (as in the Thesaurus of Geographic Names). Here we discuss feature type schemes and the various methods used to store these schemes and how this choice has, at times, dictated the structure of the gazetteer as a whole.

5.2.1 Feature Type Schemes. Most digital gazetteers developed during the end of the 20th century and the beginning of the 21st century have developed their own feature type schemes [68]. This situation has plagued digital gazetteers and left a persistent problem in gazetteer interoperability, leaving even modern gazetteer developers and users grappling with the issue of inconsistency in the use of descriptors or codes for feature type. For example, in GNIS, which uses a broad categorization binning many discrete types together, "Park" could refer to reserves, forests, monuments, lakes, historical or archaeological sites and even cemeteries and museums. On the other hand a "Park" in a scheme that is more specific, like GeoNames, means an often forested place maintained for recreation and beauty. GeoNames gives cemeteries, reserves, forests, lakes, and historical sites their own codes. This problem dates back to the beginning of authoritative digital gazetteers in an era where digitization was still new, and perhaps interoperability was not such a high priority. The UN ECOSOC[19] attempted to build a standard for gazetteers, and was mostly focused on global coverage. It was not adopted in the numerous gazetteers developed by individual governments that used their own sets of feature types, reflecting different cultures, languages and countries [133].

While online gazetteers, such as those of the U.S Geological Survey and the U.S. National Geospatial-Intelligence Agency (formerly National Imagery and Mapping Agency), Australian Geographic Names Gazetteer and the Getty Thesaurus of Geographic Names, used their own feature type schemes, the ADL FTT (Feature Type Thesaurus) was the first large scale effort to merge several of these schemes. The FTT uses types found in the schemes of its source gazetteers and place name records, which are referred to as lead-in terms that are linked to their corresponding preferred terms in the FTT [67, 68]. While manually mapping the existing schemes to their own, the ADL FTT also generates feature types using common "surnames" of places. "Surnames" refer to commonly found end positions of place names (e.g. the last word of the place name Waikato River indicates its feature type). They captured not only single word feature types that are commonly found after proper nouns of place names but also multiple word feature types like "oil seep", "gas field" etc [68]. The ADL FTT included 209 preferred terms that accommodated 978 lead-in terms from existing feature type schemes. The FTT consisted of 6 top level feature types: administrative areas, hydrographic features, manmade features, physiographic features, land parcels, and regions. These top level features were further broken into 3 more levels.

The ADL data model along with the FTT became adopted by numerous gazetteers [92, 96, 126, 151]. There were also attempts to enrich the feature types in the FTT using further extraction of generic terms, correlating them to types already in the FTT and placing them in suitable positions in the FTT. Thus [153] used commonly occurring tokens in place names that show correlations with their respective feature type as potential lead-in terms to the FTT, using hierarchical clustering to identify the correct insertion location of the FTT. The authors of [153] note that their efforts led to the identification of numerous non-English generic terms. Other gazetteers that have global coverage, notably GeoNames, have their own feature type schemes, which were in turn sometimes adopted by other gazetteers [119].

Other efforts by gazetteer developers to create their own type schemes included The Information Commons Gazetteer with a simple scheme that accommodated places from GNIS and the US National Geo-spatial Agency's GEOnet Names Server (GNS) database [94]. Adherence to a unified data model or feature type scheme was challenging for several reasons:

[19]https://ecosoc.un.org/en

- Difference in use case of gazetteers - requiring different feature types classification schemes. For example, a general purpose gazetteer might include all historical or archaeological sites under a general type but this is inadequate for a historical gazetteer that could require specialised categorization of the various historical sites.
- Disagreement on preferred names – e.g. gazetteer developers may choose to use synonyms of the terms used in the source gazetteer thesauri. Thus the ADL FTT lists lead-in terms (from its sources) that are synonymous to the terms they have chosen to represent a concept.
- Cultural biases – The descriptions and the vocabularies used in different gazetteers are culturally biased and may not suit all users from all regions and cultures. The meaning of a place referred to by a name can vary between cultures, and hence the naming of places varies with the prominent language of these regions. Another issue is the under representation of some languages, cultures and regions due to historical events and power structures.

5.2.2 Feature Type Scheme Formats. Here we discuss feature type schemes in gazetteers and their relative merits, with a focus on thesauri and ontologies. Thesauri were used in some early digital gazetteers, notably the Getty Thesaurus of Geographic Names. However, their use is accompanied by some shortcomings. In particular:

- Thesauri only supported a limited number of relations, typically of hierarchical, associative and equivalence [112].
- Most feature type thesauri did not enable multiple hierarchies, making it difficult to represent some entities as perceived by humans.
- Entities in thesauri may not have specific characteristics that govern the establishment of relations, as entities are not as robustly defined as concepts in ontologies, in which definitions of concepts can govern their relations.
- The definition of a concept as an entry in a thesaurus is informal and may be insufficient for further integration and expansion in different applications.
- Problems in interoperability such as resulting from the previous point.

It is to overcome these issues that gazetteer developers started adopting alternative feature type approaches, particularly ontologies. Interest in using ontologies (geo-ontologies) for feature types in gazetteers coincided with the emergence of the concepts of the semantic web and linked data [17, 19, 84]. The more formal nature of an ontology and the fact that it can be regarded as a shared concept among various parties [21] can be seen to help alleviate some of the issues with thesauri. A comparison between the use of thesauri and ontologies for feature type schemes is shown in Table 2.

Early attempts to build ontologies for gazetteers were based on existing gazetteer feature type schemes [75]. With ontological approaches, the boundary between an ontology as a feature type scheme and building a place name ontology that could be populated with instances of places and used as a gazetteer begins to blur [60, 95, 116, 136].

Adopting an ontology based approach for a gazetteer claims several advantages over the traditional list/taxonomy or thesaurus based approaches. In [43, 50], the authors point out that an ontology based approach can improve spatial querying with regards to vague place names such as "South of France". The ability to define arbitrary relations among concepts that can aid in querying, and reasoning is another advantage of ontologies. Relations like *hasOrigin* or *hasDestination* are examples in which a stream can have a spring as an origin, and a tributary can have a river, or a river an ocean, as its destination. [115, 137] present an ontological gazetteer that explicitly stores topological relations such as 'within' and 'touch', and vague relations such as 'near'. GeoSPARQL

Table 2. Thesaurus and ontology comparison

	Thesaurus	Ontology
Application	Information retrieval and structuring.	Inference and reasoning, information retrieval.
Enforced hierarchy	Generic, whole-part or instance hierarchy.	is-a hierarchy.
Relations supported	Restricted number of relations.	Arbitrary number of relations.
Formal semantics	No formal semantics to support reas--oning.	Formal semantics enables deductive reasoning. This allows inference of new knowledge about relations between concepts.
Definition of terms and concepts	Terms representing concepts.	Formal specifications of concepts.

[12, 28], which is an extension to SPARQL [121], by the Open Geospatial Consortium (OGC)[20], allows querying semantic geographic data on the web using 9-IM/RCC topological relations [34, 35], further improving the functionality of ontology-based approaches [85]. Improved reasoning in Geographic Information Retrieval (GIR) applications is made possible due to the expressivity of description logic associated with some ontologies – thus enabling subsumption and similarity based reasoning [75]. Further benefits of ontologies are that they can support multiple hierarchies and have the potential to be more maintainable, and scalable and to facilitate interoperability. The benefits are however offset by higher costs of implementation.

It is important to note that while some of the most widely used gazetteers today, like GeoNames, Nominatim and the Getty Thesaurus of Geographic Names may have published their data as Linked Data and built ontologies around them, they are still operating primarily with either simple feature type taxonomies or feature type thesauri.

5.3 Geographic Footprint

The geographic footprint is the quantitative geometric entity that links a toponym to its real world location. Latitude and longitude coordinate pairs in the form of points, bounding boxes, lines, polygons and grid references can be used to represent the geographic footprint of a place. For instance, in the ADL Gazetteer Content Standard (GCS), it is compulsory for each feature to have one or more geometries (points, lines or polygons) and a bounding box which is a generalisation of the geometries. It is possible for a single place to have multiple geometric representations -

- from different sources that represent the same place differently
- different footprints representing the changes of the location over time
- represented with multiple types – for example a point and a bounding box, or a point and a polygon which represent the object at different levels of generalisation
- multiple representations of linear or polygon geometry according to level of generalisation

Geometric footprints stored in gazetteers are clearly very much an approximation of the location of the represented place. Spatially extensive features such as cities, lakes, or rivers are commonly represented just by a zero-dimensional point, but almost all real world places require at least two dimensional representation for better fidelity. Lines are sometimes used to represent linear features like roads or rivers, although all such natural linear features have a width. Bounding boxes can

[20]https://www.ogc.org/

provide an indication of extent of a feature but how well they do so varies with its actual form. Thus, for some features like cities it could be quite useful but for some linear features it might grossly overestimate the actual area extent.

While a polygon with a possibly large number of vertices can be used to model the boundary of some real world features with high accuracy, natural features do not usually have crisp boundaries and therefore their representations are still inevitably approximations to some degree. Very detailed geometry can also introduce computational overheads. Due to the unavoidable imprecisions introduced into gazetteers when modelling locations, [68] has suggested that gazetteers should convey the degree of approximation.

5.3.1 Approximating the geometric extent of gazetteer features. In [157], the authors propose using circles for footprints when attempting to derive the extent of populated places with only point locations. They use the mean distance between point pairs within a particular region of a county to calculate a radius for these point buffers. Although intended to represent the fuzziness in the boundaries of populated places the method may have varying suitability depending on the density of named places and their population density, and is clearly focused on non-natural features. [134] focus on improving minimum bounding boxes (MBB). They test out hierarchical, geometrical, probabilistic and heuristic methods to represent a Minimum Bounding Rectangle (MBR) and test their methods on data from GeoNames and OSMnames[21] using Google map MBBs to verify. They argue their probabilistic approach can be used to improve footprints of gazetteer place names at district level or larger while their heuristic and geometrical approaches can be used to improve data on places that do not have enough coverage. They use the parent-child relationships given in gazetteers to model the MBR of the parent place and use optimization methods to include as many children as possible but also to reduce the distance between the point location given for the parent place and the center of the MBR they derive.

Given prior knowledge of the containment and non-containment of point-referenced places within larger regions, that could be vague vernacular places, Voronoi diagram methods were applied in [6]. In an approach to modelling the extent of sets of points contained within vague regions with known names, obtained through web search methods, several methods based on Delaunay triangulation were applied in [7]. The boundaries of vague place names were estimated in [165] based on their natural language spatial relationships to places with known coordinates from Chinese gazetteers. They generate boundaries by combining geometric applicability models (also known as spatial templates) of multiple individual natural language spatial relationships to the known places.

5.3.2 Modelling extents with kernel density estimation (KDE). One of the most common methods for representing the extent of places, based on multiple point data samples for a named place retrieved from VGI and web pages, is to apply kernel density estimation (KDE) to the points [24, 70, 72, 87, 141, 145, 147]. When using social media, a significant problem can arise with introducing bias in the inferred region [70]. Thus, individual contributors could skew the data to a particular person's perception of the location of a place. Hence when there are multiple contributors to a set of data it could be appropriate to limit the number of contributions from each user. Bulk uploads can also skew results, thus if there are many tags with identical coordinates only one such coordinate might be selected. A problem specific to KDE is deciding on the boundary of the resulting fuzzy region. Various threshold values for the volume of the KDE surface have been proposed, such as between 50% and 90%, e.g. [24, 70, 77]. Similarly different KDE bandwidth values have been used, in the case of [24] there being two, for different levels of granularity, each in combination with

[21] https://osmnames.org/

thresholds. As an alternative to KDE for modelling sets of points, fuzzy set based methods with thresholds were applied in [52].

5.3.3 Discrete global grids. Discrete global grid systems are another technique used for grounding place names (i.e. generating geometric representations for them) in gazetteers, though not many were identified in our study. Wāhi [3] is a discrete global grid gazetteer that implements three grid reference systems.

Fig. 4 illustrates some of these methods used to record geographic footprints.

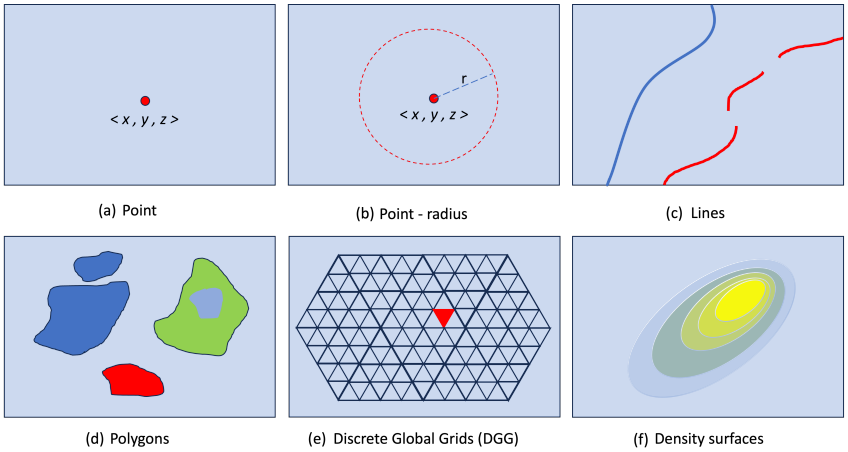

Fig. 4. Various methods of recording geographic footprints. As shown in (a), a point is defined by an 'x' and a 'y' coordinate pair. Gazetteers may also optionally store a third coordinate for the vertical height or elevation from some reference level. (b) shows a point-radius representation where a radius distance, 'r', accompanies the point. Lines can be represented as a continuous collection of points or a collection of disjoint line segments as shown in (c). (d) shows different types of polygons: a simple polygon in red, a multi-polygon in blue and a complex polygon in green which has a hole in the center. (e) shows a location in a discrete grid in red. Density surfaces produce a raster grid with each cell representing a density as shown in (f).

5.3.4 Temporal change in footprints. Some research has focused on the update of footprints to reflect the gradual change with time of features such as human settlements and forests, lakes and rivers. Satellite images [110] and aerial images [4] of these features can be used to enhance gazetteers with changes over time. [110] used satellite images with visual appearance models to estimate the bounding boxes, while [4] use a least squares matching algorithm to detect changes in boundaries of features, comparing the boundary stored in their spatio-temporal gazetteer with new images of the feature. With the abundant availability of satellite imagery datasets and advances in computer vision, automatic generation and temporal maintenance of complex footprints from these sources has considerable potential.

5.3.5 Efficiency in footprint storage. The alternative ways of representing footprints alluded to earlier are accompanied by issues of storage and computational efficiency. Not all applications need features to be delineated in high detail. In this context [68] introduces satisficing into digital gazetteer footprints, i.e. not to seek optimal solutions to certain problems because the costs are too high, but to settle for solutions that are satisfactory given the cost. With the advent of open and hence sharable linked data, the requirement of many gazetteers to store complex geographic objects comes into question, as opposed to referencing a particular geometric representation. Thus

storage of duplicate complex geometries can be avoided [93, 125], though it does not necessarily reduce the computational overhead for the end user.

5.4 Temporality

[67] identifies the need to record temporality in every major component of the gazetteer and hence achieve a truly spatio-temporal gazetteer. In the GCS, every place name, feature type, footprint, relationship, classification and other item of data associated with the place (including attributes such as population) needs to have a time period associated with it. In addition to start and end time data items, in association with labels of current, former or proposed, the GCS supports recording a detailed time period, a named time period and notes on the time period.

Temporal data can reflect change of urban boundaries, civilizations, kingdoms, rulers and governing states as well as aspects of the continuous change in the natural environment. Changes in the name of a place might affect our perception of it. It could be argued that changes in the name of city due to change in the governing regime but not to a large extent in the population, could be perceived differently from, say, the change in a school from a secular to a religious one in which the entire ethos and much of the student population changed. An example of the former is Saint Petersburg in Russia being Petrograd in the early 20th century, before becoming Leningrad in honour of the Russian Communist leader, and then reverting to Saint Petersburg with the fall of communism. The perception might change according to the prior knowledge or experience of the perceiving individual, but also in the nature of the change. In the case of the school it might be regarded as less different if only the name changed but everything else remained the same. Similar issues could arise in the change in the feature type of a place. The change of geometric footprints could also result in different perceptions, for example if a small settlement evolves to a large conurbation or a lake dries up or grows, with respective environmental impacts.

Temporality becomes a key focus when dealing with historical gazetteers. In [136], the authors point out the difficulty in recording or finding out when a place started or stopped being called by a name. They point out that even vague dates are hard to find when searching for information and argue that gazetteers should record as many historical names as possible and treat the usage of names as attestations of their validity.

Capturing temporal changes using aerial images [4] and satellite images [110] was discussed earlier under geographical footprints. In order to be temporally scoped it is not only important to update the footprints but to timestamp these footprints.

Following the ADL GCS, many gazetteer developers accommodated some sort of temporal attributes to their models [30, 42, 46, 109]. Furthermore, the latest formats for storing place records, especially on the linked web, facilitate temporal scoping. The GeoJSON-T format is an enhancement to the GeoJSON format (the T stands for temporal) and the latest Linked places format is built on top of the GeoJSON-T format [46].

5.5 Stored and Inferred Relations Between Places

One aspect of the place name component is that a name may include or be directly associated with its hierarchical parents, particularly administrative regions, such as county, state, province and nation. In addition to such specified hierarchical information, some state agencies and research projects developed gazetteers that also supported functionality to compute topological relations from footprint geometry [128, 142, 154]. Related place name knowledge resources referred to variously as ontological gazetteers [95] and place name ontologies [76, 116] stored spatial relations explicitly rather than computing on the fly. In the case of [76] the model included reference to application-specific data items (artefacts) as well as a feature types from an external feature type thesaurus.

5.6 Other

While the main components of gazetteers are those described above, some gazetteers do include other types of information that can go some way towards the vision of the richer representation of the named place that was referred to in the introduction. For example, the TGN has a category of Place Type, which while equivalent in some ways to a feature type (and includes the main feature type) is an explicit recognition of the concept of place. Its entry for the Italian city of Venice is as follows:

```
inhabited place (preferred, C)  .... settled by refugees from Lombard invasion
                                      after 568
city (C)
commune (administrative) (C)
regional capital (C)
provincial capital (C)
cultural center (C)  ............  flourished during 15th-18th cen.; home to many
                                    important Renaissance artists & architects
trade center (C)
military center (H)
republic (H)  ................  from middle of 12th cen. until 1797
World Heritage Site (C)  .........  since 1987
first level subdivision (C)
```

The ADL GCS originally included a single ”description” field to store a free text description of a place, but this was later modified to allow multiple descriptions reflecting the different facets of a place. The Pleiades gazetteer described as “A Gazetteer of Past Places”[22] also records place types which typically refer to cultural origins, such as “sanctuary (religious center)”. This gazetteer also often includes free text descriptions of the place referring to its history. Examples of other types of content include “History” and “Description” fields in the GNIS, where the latter is often used to describe the geographic context of a place. The New Zealand NZGB gazetteer[23] has “Events” and “History/origin/meaning”, and the TGN has a “Notes” field usually used to describe the historical origins as well as references to important cultural sites. The TGN Notes entry for Venice reads as follows:

The islands of the Venetian lagoon were settled by refugees from Altino and Aquileia after the Lombard invasion of northern Italy in 568. The settlement expanded into the Adriatic and Aegean from the 11th century. It was ruled by Austria, then France, 1798-1848. It went to Italy in 1866 The entire city including the lagoon was named an UNESCO World Heritage Site in 1987.

Some of the richest place related content is to be found in resources that are not described as gazetteers. One of the most widely used examples is the set of the place specific pages in Wikipedia, which are complemented by the more structured representations of DBpedia and Wikidata which include data properties such as *inception* (a date), *significant event*, *birth place* of, *death place* of, *comment*, *abstract* (a summary of the Wikipedia content), *residence* of, *garrison* of, *type* (including Yago[24] place types). Other example resources are more localised as in Kā Huru Manu (The Ngāi

[22]https://pleiades.stoa.org/places
[23]https://gazetteer.linz.govt.nz/
[24]https://yago-knowledge.org/

Table 3. Distribution of various sources that have been integrated to compile gazetteers over the years. Each number refers to a citation.

	1999-2002	2003-2006	2007-2010	2011-2014	2015-2018	2019-2022
Integrating authoritative gazetteers	[69], [31], [103]	[126], [94]	[23], [151], [73], [66], [158]	[162]	[114], [16], [3], [64], [65]	[61], [93], [117]
Digitizing ancient gazetteers/ collection records		[55], [78]	[136]	[159], [30]	[20], [58], [114], [14], [29], [64], [155], [82]	[32],[46], [42],[117], [90]
Digitizing maps			[106]		[64], [160], [86]	[160], [64], [32], [71], [89]
Integrating VGI and Authoritative sources			[147], [148], [96], [118], [135]	[13], [161], [140], [108]	[120], [125]	
VGI sources		[167]	[124], [5], [77], [119], [56], [118], [81], [115]	[22], [127], [54], [83]	[41], [120], [113], [52], [101]	[72]

Publication Date

Tahu Cultural Mapping Project)[25] from New Zealand which focuses on the Māori origins and stories associated with named places.

6 Gazetteer Sources and Integration

In this section, we explore the sources used to compile modern gazetteers and methods used in their integration. Most gazetteers and papers that were retrieved in our search, produced in the late 20th century or 21st century, were compiled by integrating other gazetteers. All modern digital gazetteers retrieved from our search fall into one of the following types depending on their sources:

(1) Gazetteers formed by the digitization of (and integration of one or more) historical gazetteers.
(2) Gazetteers formed by toponym extraction from existing maps either modern or historical.
(3) Gazetteers formed by integrating other digital gazetteers.
(4) Gazetteers formed by integrating VGI (or web harvested information) with existing gazetteers.
(5) Gazetteers formed completely from VGI or web harvested data.

Table 3 shows the distribution of research papers in our review over time that used each of these sources, and indicates that most of the recent gazetteers were created by integrating or improving one or more already existing authoritative or VGI digital gazetteers.

[25]https://kahurumanu.co.nz

As evident throughout the literature, all the methods listed above excluding (2) and (5) have gazetteer integration as a key element in gazetteer compilation. Gazetteer integration in general requires identification of duplicates between entries in the gazetteers and then resolution of differences to combine them as a single record. Both these tasks are complicated by the existence of multiple names for the same place and different feature type schemas (potentially resulting in different feature types for the same named place), and the variety of geographic footprints available along with their source inaccuracies.

Duplicate identification or entity resolution, also known as gazetteer matching in this context, is the process of matching features (gazetteer records) that refer to the same real world place, either between gazetteers or within the same gazetteer. Methods for resolution of duplicates often use a combination of string similarity between toponyms, geographical distances between the features (or other forms of geometric relationship like overlap) [129, 166] and a semantic distance based on the feature types. In this section we summarise some of the methods used in the literature. Note that nearly all of the cited papers refers to gazetteers. There is however a body of literature that applies similar methods to match points of interest (POI). We have cited a couple of recent such papers as they employ techniques that are relevant to gazetteers [11, 74].

The duplicate identification task can be classed as either intra-gazetteer or inter-gazetteer. Inter-gazetteer matching is considerably more complex due to the challenges in matching different feature type schemas [69]. Table 5 (Appendix) presents a summary of publications that either present methods for gazetteer integration or present gazetteers that have been compiled by integrating other gazetteers and that discuss deduplication techniques. This table has columns indicating whether integration is inter- or intra-gazetteer or both; the data sources; the main types of information used for matching records with regard to the use of the names, footprints and feature types; and notes on how the features are exploited focusing on the types of classifier or other statistical methods employed in machine learning techniques. In column Feature Name we list the techniques for name matching applied in cited papers. These mostly refer to string similarity measures such as Levenshtein edit distance, and Jaro-Winker, Jaccard, Monge-Elkan, and Soundex similarity and distance measures. Geographic footprint similarity measures are based on distance between the footprints, the degree of overlap, or containment within other footprints. The Feature Type column refers to the use of explicit feature categories and any text that might indicate aspects of the feature type (e.g. 'eating'). Here similarity measures include distance between terms in a class hierarchy, distance to a common parent in a class hierarchy, presence or absence of the type in a vector of all types (i.e. one-hot encoding), as well as Dice coefficient, and the Jaccard and the Wu and Palmer distance between the type terms. Feature type terms, and feature names, can also be represented as word embeddings when input to machine learning classifiers.

The various measures of similarity have either been used with explicit thresholds in rule-based methods, or as input features for a machine learning classifier. In the case of more recent deep learning transformer-based methods [11] similarity is learnt from the word embeddings in the training data, without use of explicit string similarity metrics.

Most of the papers listed in Table 5 (Appendix) refer to applications of integration techniques that were applied when building a gazetteer, but these techniques could also be applied in a more dynamic context in which a gazetteer service provides online access to multiple external sources that are matched and integrated at the time of user query. This approach has been referred to as a meta-gazetteer [135] and has similarities with for example the extraction, transformation and loading (ETL) approach of [96].

Except for the techniques in [33, 96, 108], the methods summarized in Table 5 (Appendix) are applicable for instance level matching, independent of the underlying KOS. However, schema level gazetteer alignment or the alignment of the underlying data structures is a broader research topic.

The alignment of two gazetteer KOS creates a new KOS with a new typology of features that match and retain the typology or concepts of the original sources to varying extents. Generic ontology alignment is well researched [8, 132]. However, standards and frameworks for place ontology alignment are less common and rarely replicated.

Generically, ontology alignment can be broadly categorized as concept level alignment or structure level alignment [132]. During alignment, a concept or element level alignment compares labels and definitions of concepts or terms in ontologies or thesauri respectively. This approach aims to exploit lexical or semantic similarities of the concepts. Conversely, structure level alignment considers concepts' relations to other concepts and their relative placement within a KOS (e.g., within a graph or a tree structure). We note that most of the gazetteer thesauri or geospatial ontology alignment techniques are dominated by concept level alignment approaches. Therefore, it is more apt to categorize the gazetteer ontology and thesauri alignment depending on whether the methods used are manual or automatic (see below).

Gazetteer data structures are manually aligned often during the creation of a new gazetteer [33, 60, 95, 96, 116, 125]. This process is time intensive, requires domain expertise and is not generalizable. Inspired by [149], Janowicz and Keßler propose a framework that can be used to align a thesaurus with an ontology [75]. The application of this framework is demonstrated by creating a place ontology using the ADL FTT. [85] presents a re-usable framework for multilingual place ontology alignment based on homologous relationships between entity toponyms, footprints, and types. They take a bottom up approach where they match the instances first in order to align the underlying ontology.

Entity matching techniques detailed in Table 5 (appendix) can be used for automatic gazetteer alignment in a bottom up fashion where matches between entities or instances are used to integrate the broader concepts or feature types they belong to. [23] manually annotate a dataset of matching features that are then used to integrate two thesauri based on the statistical matching frequency between feature type pairs. Similarly using annotated data, methods described in [2, 11, 66, 98, 99] use machine learning algorithms to align gazetteer data structures. OWL:sameAs[26] links have also been exploited to align gazetteers [95, 110]. [139] is an example of a structural approach to aligning place ontologies where the authors use a combination of concept level similarities as well as inheritance relations and relations between siblings. These automatic alignment methods are more generalizable and also require less domain expertise. When published as Linked Open Data, the methods can also lead to significant reduction of data redundancy, though they are often hindered by the absence of relations such as OWL:sameAs and SKOS:exactmatch[27] in ontology implementations [40].

7 Volunteered Geographic Information

The phrase Volunteered Geographic Information (VGI) was adopted in [57] to refer to shared geographic information that was user generated, or crowdsourced, though crowdsourced geographic information was by then quite well established, for example with Wikimapia. Introduction of the term VGI coincided with the idea of non-authoritative, geographic knowledge systems that were self-governed by the public [18]. VGI has become a major topic in the building of all geographic knowledge systems including gazetteers, and its presence permeates almost all aspects of gazetteer building, content, and maintenance. While some gazetteers are purely VGI-based there has been interest in the enrichment of authoritative gazetteers with VGI. Such efforts are however inevitably limited by the fact that the resulting gazetteer might no longer be classed as authoritative, and

[26] https://www.w3.org/2001/sw/wiki/SameAs

[27] https://www.w3.org/2009/08/skos-reference/skos.html

Table 4. Gazetteers created using intrinsic and extrinsic gazetteers

	Geography	
	Explicit	**Implicit**
Explicitly Volunteered	**OSM** [13, 54, 83] **Panoramio** [118] **Wikimpaia** [54, 83]	**Wikipedia** [52, 118, 124] **Dbpedia** [83, 108] **Flickr** [5, 52, 118, 124]
Implicitly Volunteered	**Twitter** [41]	**Gumtree** [147, 148] **Phonebooks** [56] **Craigslist** [72] **Search Engine results** [77, 118, 120]

brings with it challenges of trust and data quality. There have however been numerous attempts to create new gazetteers integrating the two sources and incorporating the merits of both, like OSM and GeoNames.

[36] propose a typology of VGI based on the way the information was made available by the community and the nature of the geographic information. If the data was made publicly available for a specific purpose it is treated as "explicit" otherwise it is "implicit". Similarly, if the information shared was explicitly geographic in nature, it is "explicit" and if the information is not about place but geographic information can be inferred, it is "implicit". Table 4 summarises gazetteers that use VGI in Craglia's VGI typology [36].

Volunteered geographic information has great potential over authoritative sources in some contexts but is lacking in others. These advantages have been discussed in great detail in the literature [38, 48, 51, 91, 102, 127, 163] and can be summarised in the following three main points.

- VGI-based gazetteers are more accessible than authoritative gazetteers. VGI is almost always free and open source, not limited by restrictive licences of some authoritative gazetteers that require payments. This accessibility of VGI makes it attractive for students and researchers.
- VGI can be more up-to-date than authoritative sources. Updating a large authoritative knowledgebase is potentially a substantial task, often rolled out as a batch of small changes. In contrast, local changes in places may be noticed by individuals who can themselves update the VGI resource, where the changes might be regulated by other participants in the same area. The ability to make timely gazetteer (and associated map) updates at short notice is extremely helpful during disasters, requiring quick responses and situation awareness.
- VGI can reflect local knowledge of individuals familiar with the locale, as opposed to being governed by a central body of experts with limited knowledge of local places. This enables VGI to capture vernacular and vague place names and identify their footprints with contextual information. Local knowledge can also give a high level of detail to features absent in authoritative sources.

7.1 Quality and Trust

Drawing from [18], the authors of [80] focus on the challenges of integrating VGI in gazetteers and present a framework for trust in VGI. They argue that trust can be used as a proxy for the quality of VGI data and present a set of basic requirements for a trust model:

(1) Minimal metadata requirements from the user's end.
(2) A feedback system must be in place that enables users to rate or assert the trust in the data obtained from the gazetteer.

(3) Computational models must be built to capture and compute a user's reputation – a metric the authors define as the collective opinion of other system users on a particular user.
(4) The trust model must be transparent allowing users to learn how someone becomes trustworthy.
(5) Provenance must be captured and integrated into the trust model – while identifying the difficulty in capturing provenance in VGI, authors suggest recording alternative data such as user's interaction history with the system, trust ratings, cumulative user reputations, time stamps of contributions or modifications and user profile information.

Several indicators have been explored in the literature for measuring the quality of VGI:

- Credibility: credibility can be determined from the source of the piece of information and, like provenance, can be hard to capture in VGI. Expertise of the user is another measure of credibility [49].
- Local familiarity: The contributor's familiarity with the locale can be used as an indication of their expertise in local knowledge [130].
- Experience: The user's experience of the platform can be integrated in models that calculate the user's reputation as discussed above [150]. It can be reflected for example in the amount of time spent using the platform, the volume of content created and or used, and the number of logins.
- Reputation: The recognition of the user among other users. Indirect methods of calculating reputation would be to use past interactions of the user with other users, and the use of datasets published by the user. A rating system can be used to explicitly capture a user's standing among peers [100].

Apart from the challenge of quality and accuracy of the data provided, VGI can also suffer from other disadvantages like sparsity and uneven coverage. The volume of volunteered data in VGI gazetteers has been seen to vary from continent to continent and country to country [1]. While population sizes of countries can be seen as an obvious reason for this disparity, [59] argue that other factors like a country's population's access to the internet and the state's policies on open data can also have a significant impact on the coverage. [1] also point out correlations between various feature types with the volume of information available – populated places had a strong correlation with feature counts when compared to natural feature types like mountains, hills and streams.

Information directly obtained from VGI platforms or scraped from implicit crowd platforms contain most of the same components (geometry, feature type, temporality etc, see Section 5), required to construct or contribute to a gazetteer, and face the same challenges that features from authoritative gazetteers encounter. Presentation of specific techniques to improve these VGI components is out of the scope of this paper. Interested readers can refer to the review of these quality measures and improvements in [130].

8 Gazetteer Technologies

Early digital gazetteers that were digitizations of printed gazetteers were easily accommodated in databases such as relational database management systems (RDBMS). Earlier versions of the GNIS (Geographic Names Information System), Ordnance Survey gazetteer of the United Kingdom and the Alexandria Digital Library gazetteer are examples of gazetteers that used RDBMS. Subsequent gazetteer technologies particularly in the last couple of decades have been dominated by the use either of extended (or object) relational databases, with their rich support for spatial data management, or of linked open data (LOD). Interest in the use of LOD was spurred by the appeal of

creating openly accessible shared content that could be managed and queried using relatively standardised internet-based technologies. LOD uses semantic web technologies [84] that are designed to provide machine-readable content on the internet [17, 19]. Linked open data is implemented with RDF (Resource Description Framework) that links entities or things identified by URIs (Uniform Resource Identifiers). RDF represents information as a set of triples, consisting of a subject, predicate, and object.

[80] provide a possible stack of technologies that can be used to implement a distributed gazetteer. Even though the focus of that paper is mostly on integrating and building trust around VGI, their suggestion of using RDF triple stores with the SPARQL Protocol and RDF Query language (SPARQL) was reflected in later gazetteers that used these technologies [64, 65]. SPARQL is the standard query language for RDF data stores and is used to query and manipulate data. Analogously with the development of SQL, it was extended to support geo-spatial data access in GeoSPARQL [12, 28].

FODGS [115] uses RDF and SPARQL with JENA-TDB, a scalable open source triplestore database by Apache. Titan, another open source distributed graph database management system that is built on top of Apache Cassandra, can be used as an RDF datastore [108]. Unlike JENA-TDB, Titan does not natively support SPARQL, but uses Gremlin as its query language.

There are several examples of gazetteers that have used OWL (Web Ontology Language) to build an ontology for the features, stored as RDF triplets, and queried with SPARQL/GeoSPARQL, [13, 29, 30]. OWL is one of the most widely used knowledge representation languages for representing and sharing ontologies on the World Wide Web [9, 37, 53] and has gained W3C recommendation as a part of the Semantic Web technologies [104]. Building on top of RDF and XML, OWL adds more vocabulary for describing properties and classes. It also adds relations between classes and richer typing of properties and characteristics. It should be noted that most of these gazetteers are designed to be available on the semantic web as linked open data and thus have interoperability as a top priority.

The snippet from Fig. 5 represents a simple RDF graph that gives information about the Eiffel Tower. The namespaces are established in the first four lines. It then describes the RDF resource at http://example.com/places/EiffelTower. This is of type "Place". It uses various tags from different namespaces to give this resource properties like name, description, and the location as a latitude and longitude pair. In the third section, the OWL class with the URI http://www.example.com/ontology/Place is described. Place is described as a subclass of a Thing and a restriction is imposed on the "name" property making the minimum cardinality for a place name to be 1.

Several advantages of storing gazetteer records as Linked Open Data can be identified:

(1) Interoperability: Linked open data can assist with integration and interoperability between different gazetteers and other datasets [30, 64].
(2) Data Quality: By using open data standards and providing access to the data through a SPARQL endpoint, it is easier for other organisations and individuals to improve the data.
(3) Ease of Access: Linked open data makes it easier for end-users to access and query the data through tools such as semantic web browsers.
(4) Reusability: Linked open data allows for easy reuse of the data in different applications and domains, which can lead to new insights.
(5) Advanced Semantic searches: complex semantic searches can be performed when coupled with a powerful ontology (e.g. – OWL) [75].

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:geo="http://www.w3.org/2003/01/geo/wgs84_pos#"
         xmlns:foaf="http://xmlns.com/foaf/0.1/"
         xmlns:dc="http://purl.org/dc/elements/1.1/">

<rdf:Description rdf:about="http://example.com/places/EiffelTower">
  <rdf:type rdf:resource="http://www.example.com/ontology/Place"/>
  <foaf:name>Eiffel Tower</foaf:name>
  <dc:description>An iconic tower and tourist destination in Paris, France.</dc:description>
  <geo:lat>48.8584</geo:lat>
  <geo:long>2.2945</geo:long>
</rdf:Description>

<owl:Class rdf:about="http://www.example.com/ontology/Place">
  <rdfs:label>Place</rdfs:label>
  <rdfs:subClassOf rdf:resource="http://www.w3.org/2002/07/owl#Thing"/>
  <rdfs:comment>A class for representing physical places.</rdfs:comment>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="http://xmlns.com/foaf/0.1/name"/>
      <owl:minCardinality rdf:datatype="http://www.w3.org/2001/XMLSchema#nonNegativeInteger">1
      </owl:minCardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>
```

Fig. 5. Code snippet showing an example of a description of a place and its attributes and relations using RDF and OWL.

A recent effort aimed at building a common format for representing places is the Linked Places Format (LPF) developed jointly by the World Historical Gazetteer (WHG) project [61] and the Pelagios project [152]. The WHG uses the PostgreSQL/PostGIS relational database together with an Elastic search index for data storage. Elastic Search is an open-source search and analytics engine designed for real-time search functionality that enables full-text search. The LPF is built around JSON-LD (a format for structuring and representing linked data using JavaScript Object Notation) making it valid RDF and also valid GeoJSON (a JSON format for geographical features). They have enhanced the format enabling its recorded places to be scoped temporally with a "when" element through the use of GeoJSON-T [46]. LPF does not intend to become the unified format or data model for records of places, but to be a format that facilitates linking between gazetteers. It helps record metadata that allows users to search across gazetteers, disambiguate and identify places and annotate data with stable URIs.

PostGIS is an Object-Relational Database Management system that extends the open source PostgreSQL with support for geographic data storage and spatial query and is used in several gazetteers [3, 46, 92, 127]. Its adoption may be attributed to the support provided for many spatial data formats such as Shapefile, KML, GeoJSON, WKT, as well as a wide range of spatial reference (coordinate) systems and transformations between them. It is also a natural development of earlier very widely used and familiar relational database technology. Fig. 6 shows some of these discussed technologies in a gazetteer building application.

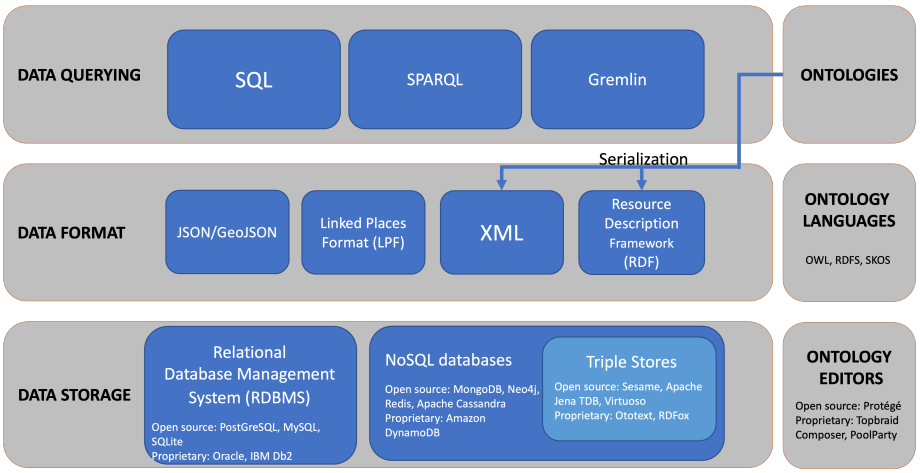

Fig. 6. Technology stack used to build modern digital gazetteers with some examples.

An issue that has been associated with the use of semantic web technologies for gazetteers is that many of the published projects are associated with academic research projects and have not been complemented with support for longer term maintenance [125].

As evident in the literature, the technologies used depend on the format of the published gazetteer. While most established gazetteers like GeoNames and OpenStreetMap have an LOD endpoint (both of them provide SPARQL endpoints), the question of what format to take depends on the use case of the gazetteer. If the gazetteer is expected to be used by expert users and is to be integrated, or interact, with other sources, or semantic searching is important, a linked data format is desirable. The downside of a LOD approach is the complexity of the task [152] and the need to learn the technologies, which might be justified if the developers of the gazetteer are planning to maintain the gazetteer and the gazetteer is not a "means to an end but an end in itself" [117].

9 Future Directions in Gazetteer Research

In conducting this review several themes have emerged that might be the subject of future gazetteer research and development. These relate particularly to their level of information richness; multi temporality and the evolution of the recorded places; integration of the multiple representations of place in different gazetteers; trust and data quality; technologies and interoperability.

We have referred to the fact that the earliest, non-digital, gazetteers consisted of descriptions of places. While there can be many perspectives on the nature of a place, digital gazetteers have the potential to play a more effective role in providing more information rich records of place names that could be used to improve support for applications in information retrieval, particularly search for places with particular characteristics. A major limitation of most digital gazetteers is their failure to record much information about the nature of the place that is named, both its physical and social characteristics and, explicitly or implicitly, what are its affordances or uses. Often this does not extend beyond a single feature type. As was highlighted in the introduction there are many types of information that could be recorded about a named place. Analogously to, but going beyond, what is done in DBpedia and Wikipedia records of named places, it would be possible to introduce a set of properties relating for example to historical origin, political governance, commercial services (including for example eating and drinking places, banks, cinemas), industries, religious institutions, cultural institutions (museums, art galleries, theatres etc), natural features,

parks, biodiversity, sports facilities, architectural styles, ethnic groups, notable inhabitants (and their dates), noise levels and crime statistics. It would also be possible to record personal experiences and perceptions based on volunteered information, as well as links to records in textual, audio and video documents relating to the place, including local and indigenous peoples' stories. The types of these data properties could be standardised to provide uniformity across gazetteers relating to different regions.

Future gazetteers could be much richer with regard to the various formal and vernacular names in different languages, in association with their temporal and spatial extents. Given that spatial extent and location can change continuously for settlements and some natural features such as rivers and lakes, there is a role for functionality to retrieve these data for a point in time or to support spatio-temporal visualisation of the change over time of named places. Such functionality might require interacting with map databases and related historical data sources. It brings with it fundamental challenges of tracking the identity of a place which could change in association with factors such as the administrative regime, population level, ethnicity, industry, commerce, architecture, cultural activity and the natural environment. With the widespread availability of satellite imagery, latest developments in computer vision can also be incorporated more robustly, especially in the change of boundaries of geographic footprints.

The current lack of consistency in the type and form of recorded information, in particular the use of feature type descriptors, or feature codes, hinders integration of multiple gazetteer sources that might refer at least partially to the same places. While it is reasonable to envisage encouraging the use of well-documented feature coding schemes, that cannot alleviate the need for methods to resolve multiple representations of the same place. The development of more accurate matching procedures is likely to be helped by the application of deep learning methods that encode words with embeddings that reflect their semantics and support determination of similarity between descriptors that could be lexically very different. In those respects, so-called blocking methods for filtering record matches have sometimes previously been quite inappropriate in using string similarity measures that could, for example, filter out equivalent names in different languages. Effective matching procedures would support the development of gazetteer portals - analogous to the meta-gazetteer concept - in which gazetteer records for a selected location could be retrieved and merged, or conflated, from multiple data sources, where those sources might differ in the types and richness of data recorded for a particular place. Instance level entity resolution methods, such as ones discussed in Section 6 under de-duplication, are potent tools in matching multiple representations of the same place. Standardisation of these problems with standard datasets and metrics (as is the case with generic entity resolution) would assist greatly in enabling them to be compared with each other. Another requirement for solving this issue is the need for openly accessible datasets, as most of the latest methods rely on machine learning methods that are often supervised techniques that require training data. Though some of the methods in Table 5 (Appendix) were papers specifically publishing methods for entity resolution, only two of them made their datasets publicly available for replication or comparison. The challenge of representing or embedding heterogeneous geometry types (e.g. a point and a polygon or a line and a multi-polygon) in a deep-learning (neural) framework also remains unaddressed in methods published to date, as most deep-learning methods rely solely on point-point distance measures to compare footprints.

The common use of volunteered data as a gazetteer source is accompanied by lack of consistency and monitoring of data quality and hence trust. While the topic of VGI data quality has received considerable attention there remains scope for development of better methods to monitor and maintain it in gazetteers, for which it is possible to envisage training machine learning methods to infer measures of quality based on the associated metadata.

There is a notable diversity in gazetteer data storage and management systems, exemplified by the contrast between relational databases and linked open data stores. Arguably attempts to try to dictate one type of data management technology are futile, but it could be beneficial to provide more widespread adoption of standard access interfaces that could be independent of the underlying storage technology. Progress has been made in this respect in the provision of APIs but there could still be benefit in providing standard forms of query as promoted by the Open Geospatial Consortium.

10 Conclusions

In reviewing the role of gazetteers as repositories of place name knowledge, we addressed the six research questions listed in the Introduction. (1) We described the evolution of gazetteers from historically recording the nature of individual places to storing a minimum of a few key components of name, geometric representation and feature type. (2) The source data of earlier digital gazetteers were often closely linked to the named content of digital maps, whether current or historical, while the content of some gazetteers has become dominated by volunteered content. Other gazetteers integrate names from authoritative map series with volunteered data while others are the result of merging some existing gazetteers. (3) When integrating gazetteer sources, one of the main challenges is that of determining whether two gazetteer records refer to the same real-world place. Automated methods for doing that have employed items of evidence based largely on the similarity of respectively, the place names, the feature types and the geometric footprints. Earlier automated methods were rule-based but current approaches usually employ machine learning including deep learning that transform the data items to be matched into word embeddings, or sometimes also geometric, embeddings. (4) We have described the increasingly important role of volunteered geographic information which in some cases dominates the content of a gazetteer. Its use is however accompanied by challenges of measuring and maintaining data quality. (5) When studying the implementation of gazetteers there is a clear distinction between those that are managed in spatially-enabled relational databases and those that are based on linked data technologies. At the time of writing it appears that relational databases are the dominant technology, as in widely used systems such as GeoNames and OpenStreetMap (with its gazetteer Nominatim) and for many national mapping agencies. Linked data technology is also well established to manage some gazetteer data both natively and to publish gazetteers implemented with other technologies. (6) We have identified significant limitations in digital gazetteers with regard to the lack of consensus on how data are represented, posing challenges in matching records between disparate sources. This relates particularly to feature types of which there are multiple classification and coding schemes. Word and geometry embeddings have potential to assist in harmonising representation. Development of improved data integration methods would benefit from much wider availability of data collections for training and testing. A further prominent shortcoming is the sparsity of information about the nature of the named place. Much richer representation of place would assist in place-based information retrieval and cultural studies. We also highlighted that knowledge-based gazetteers could record the spatio-temporal evolution of named places assisting in their role in historical information retrieval.

In summary, gazetteers continue to play an essential role as place name knowledge resources that are of value in their own right but very importantly can be used to support information retrieval tasks that require the user to specify one or more place names. There is great potential for future research and development in the field, particularly with regard to providing much richer repositories of knowledge of places and of the ways in which named places have evolved over time and space and have changed in their physical, socio-economic and cultural nature.

Acknowledgments

The research was funded by the New Zealand Ministry of Business, Innovation and Employment (MBIE), grant number MAUX2104.

References

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

[2] Elise Acheson, Michele Volpi, and Ross S Purves. 2020. Machine learning for cross-gazetteer matching of natural features. *International Journal of Geographical Information Science* 34, 4 (2020), 708–734.

[3] Benjamin Adams. 2017. Wāhi, a discrete global grid gazetteer built using linked open data. *International journal of digital earth* 10, 5 (2017), 490–503.

[4] Peggy Agouris, Kate Beard, Georgios Mountrakis, Anthony Stefanidis, et al. 2000. Capturing and Modeling Geographic Object Change: A Spatio Temporal Gazetteer Framework. *Photogrammetric Engineering and Remote Sensing* 66, 10 (2000), 1241–1250.

[5] Shane Ahern, Mor Naaman, Rahul Nair, and Jeannie Hui-I Yang. 2007. World explorer: visualizing aggregate data from unstructured text in geo-referenced collections. In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*. 1–10.

[6] Harith Alani, Christopher B Jones, and Douglas Tudhope. 2001. Voronoi-based region approximation for geographical information retrieval with gazetteers. *International Journal of Geographical Information Science* 15, 4 (2001), 287–306.

[7] Avi Arampatzis, Marc Van Kreveld, Iris Reinbacher, Christopher B Jones, Subodh Vaid, Paul Clough, Hideo Joho, and Mark Sanderson. 2006. Web-based delineation of imprecise regions. *Computers, Environment and Urban Systems* 30, 4 (2006), 436–459.

[8] Fatima Ardjani, Djelloul Bouchiha, and Mimoun Malki. 2015. Ontology-alignment techniques: survey and analysis. *International Journal of Modern Education and Computer Science* 7, 11 (2015), 67.

[9] Helbert Arenas, Benjamin Harbelot, and Christophe Cruz. 2013. A semantic web approach for geodata discovery. In *International Conference on Conceptual Modeling*. Springer, 117–126.

[10] Andrea Ballatore, David C Wilson, and Michela Bertolotto. 2013. A survey of volunteered open geo-knowledge bases in the semantic web. In *Quality issues in the management of web information*. Springer, 93–120.

[11] Pasquale Balsebre, Dezhong Yao, Gao Cong, and Zhen Hai. 2022. Geospatial entity resolution. In *Proceedings of the ACM Web Conference 2022*. 3061–3070.

[12] Robert Battle and Dave Kolas. 2012. Enabling the geospatial semantic web with parliament and geosparql. *Semantic Web* 3, 4 (2012), 355–370.

[13] Kate Beard. 2012. A semantic web based gazetteer model for VGI. In *Proceedings of the 1st ACM SIGSPATIAL international workshop on crowdsourced and volunteered geographic information*. 54–61.

[14] Mafkereseb Kassahun Bekele, Rolf A De By, and Gaurav Singh. 2016. Spatiotemporal information extraction from a historic expedition gazetteer. *ISPRS international journal of geo-information* 5, 12 (2016), 221.

[15] J Lennart Berggren and Alexander Jones. 2000. *Ptolemy's Geography: an annotated translation of the theoretical chapters*. Princeton University Press.

[16] Merrick Lex Berman, Johan Åhlfeldt, and Marc Wick. 2016. Historical Gazetteer System Integration: CHGIS, Regnum Francorum, and GeoNames. In *Placing Names: Enriching and Integrating Gazetteers*. Indiana University Press, 110–126. http://www.jstor.org/stable/j.ctt2005zq7.13

[17] T Berners-Lee. 2006. Linked Data. https://www.w3.org/DesignIssues/LinkedData.html

[18] Mohamed Bishr and Werner Kuhn. 2007. Geospatial information bottom-up: A matter of trust and semantics. *The European information society: Leading the way with geo-information* (2007), 365–387.

[19] Christian Bizer, Tom Heath, and Tim Berners-Lee. 2011. Linked data: The story so far. In *Semantic services, interoperability and web applications: emerging concepts*. IGI global, 205–227.

[20] Daniel Blank and Andreas Henrich. 2015. Geocoding place names from historic route descriptions. In *Proceedings of the 9th Workshop on Geographic Information Retrieval*. 1–2.

[21] W.N. Borst. 1997. *Construction of Engineering Ontologies for Knowledge Sharing and Reuse*. PhD Thesis - Research UT, graduation UT. University of Twente, Netherlands.

[22] Alessio Bosca and Luca Dini. 2009. Automatic gazetteer generation from Wikipedia. In *Natural Language Processing for Digital Libraries Workshop*. Springer, 61–71.

[23] Daniela F Brauner, Marco A Casanova, and Ruy L Milidiú. 2007. Towards gazetteer integration through an instance-based thesauri mapping approach. In *Advances in Geoinformatics: VIII Brazilian Symposium on GeoInformatics, GEOINFO 2006, Campos do Jordão (SP), Brazil, November 19–22, 2006*. Springer, 235–245.

[24] Paul Brindley, James Goulding, and Max L Wilson. 2018. Generating vague neighbourhoods through data mining of passive web data. *International Journal of Geographical Information Science* 32, 3 (2018), 498–523.
[25] Michael Buckland, Aitao Chen, Fredric C Gey, Ray R Larson, Ruth Mostern, and Vivien Petras. 2007. Geographic search: catalogs, gazetteers, and maps. *College & Research Libraries* 68, 5 (2007), 376–387.
[26] Niclas Burenhult and Stephen C Levinson. 2008. Language and landscape: a cross-linguistic perspective. *Language sciences* 30, 2-3 (2008), 135–150.
[27] Guoray Cai. 2002. GeoVSM: An integrated retrieval model for geographic information. In *International Conference on Geographic Information Science*. Springer, 65–79.
[28] Nicholas J Car and Timo Homburg. 2022. GeoSPARQL 1.1: Motivations, details and applications of the decadal update to the most important geospatial LOD standard. *ISPRS International Journal of Geo-Information* 11, 2 (2022), 117.
[29] Silvio D Cardoso, Flor K Amanqui, Kleberson JA Serique, José LC dos Santos, and Dilvan A Moreira. 2016. SWI: A semantic web interactive gazetteer to support linked open data. *Future Generation Computer Systems* 54 (2016), 389–398.
[30] Silvio D Cardoso, Kleberson J Serique, Flor K Amanqui, JL Campos Dos Santos, and Dilvan A Moreira. 2014. A gazetteer for biodiversity data as a linked open data solution. In *2014 IEEE 23rd International WETICE Conference*. IEEE, 435–440.
[31] Roberto Cervellati, Chiara Ramorino, Jörn Sievers, Janet Thomson, and Drew Clarke. 2000. A composite gazetteer of Antarctica. *Polar Record* 36, 198 (2000), 278–285.
[32] Shih-Pei Chen, Kenneth J Hammond, Anne Gerritsen, Shellen Wu, and Jiajing Zhang. 2020. Local gazetteers research tools: Overview and research application. *Journal of Chinese History* 4, 2 (2020), 544–558.
[33] Gang Cheng, Xiaoping Lu, Xiaosan Ge, Haiyang Yu, Yupeng Wang, and Xiaotian Ge. 2010. Data fusion method for digital gazetteer. In *2010 18th International Conference on Geoinformatics*. IEEE, 1–4.
[34] Eliseo Clementini and Anthony G Cohn. 2014. RCC*-9 and CBM. In *International Conference on Geographic Information Science*. Springer, 349–365.
[35] Eliseo Clementini, Jayant Sharma, and Max J Egenhofer. 1994. Modelling topological spatial relations: Strategies for query processing. *Computers & graphics* 18, 6 (1994), 815–822.
[36] Max Craglia, Frank Ostermann, and Laura Spinsanti. 2012. Digital Earth from vision to practice: making sense of citizen-generated content. *International Journal of Digital Earth* 5, 5 (2012), 398–416.
[37] Huamin Dang, Jing Zhang, Dapeng Zhang, and Tun Wang. 2011. Construction of Beijing place-name ontology based on spatial cognition. In *2011 19th International Conference on Geoinformatics*. IEEE, 1–6.
[38] Nyangweso Daniel and Gede Mátyás. 2023. Citizen science characterization of meanings of toponyms of Kenya: a shared heritage. *GeoJournal* 88, 1 (2023), 767–788.
[39] Clare Davies, Ian Holt, Jenny Green, Jenny Harding, and Lucy Diamond. 2009. User needs and implications for modelling vague named places. *Spatial Cognition & Computation* 9, 3 (2009), 174–194.
[40] THVM de Moura and CA Davis Jr. [n. d.]. Linked geospatial data: challenges and research opportunities. In *Proceedings of the XIV Brazilian Symposium on Geoinformatics*. 13–18.
[41] Maxwell Guimaraes de Oliveira, Cláudio EC Campelo, Cláudio de Souza Baptista, and Michela Bertolotto. 2015. Leveraging VGI for gazetteer enrichment: A case study for geoparsing twitter messages. In *Web and Wireless Geographical Information Systems: 14th International Symposium, W2GIS 2015, Grenoble, France, May 21-22, 2015, Proceedings 14*. Springer, 20–36.
[42] Annamaria De Santis, Matteo Gallo, Irene Rossi, and Jérémie Schiettecatte. 2021. The digital Gazetteer of Ancient Arabia. *Umanistica digitale* 11 (2021), 125–143.
[43] Tiago M Delboni, Karla AV Borges, Alberto HF Laender, and Clodoveu A Davis Jr. 2007. Semantic expansion of geographic web queries based on natural language positioning expressions. *Transactions in GIS* 11, 3 (2007), 377–397.
[44] Grant DeLozier, Jason Baldridge, and Loretta London. 2015. Gazetteer-independent toponym resolution using geographic word profiles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 29.
[45] Curdin Derungs, Flurina Wartmann, Ross S Purves, and David M Mark. 2013. The meanings of the generic parts of toponyms: use and limitations of gazetteers in studies of landscape terms. In *Spatial Information Theory: 11th International Conference, COSIT 2013, Scarborough, UK, September 2-6, 2013. Proceedings 11*. Springer, 261–278.
[46] Vincent Ducatteeuw. 2021. Developing an Urban Gazetteer: A Semantic Web Database for Humanities Data. In *Proceedings of the 5th ACM SIGSPATIAL International Workshop on Geospatial Humanities*. 36–39.
[47] Max J Egenhofer. 2002. Toward the semantic geospatial web. In *Proceedings of the 10th ACM international symposium on Advances in geographic information systems*. 1–4.
[48] Randa El Khatib. 2019. Laying the Foundation for Community-Driven, Open Cultural Gazetteers. *KULA* 3, 1 (2019), 1–5.

[49] Andrew J Flanagin and Miriam J Metzger. 2008. The credibility of volunteered geographic information. *GeoJournal* 72 (2008), 137–148.
[50] Gaihua Fu, Christopher B Jones, and Alia I Abdelmoty. 2005. Ontology-based spatial query expansion in information retrieval. In *On the Move to Meaningful Internet Systems 2005: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2005, Agia Napa, Cyprus, October 31-November 4, 2005, Proceedings Part II*. Springer, 1466–1482.
[51] Georg Fuchs, Natalia Andrienko, Gennady Andrienko, Sebastian Bothe, and Hendrik Stange. 2013. Tracing the German centennial flood in the stream of tweets: first lessons learned. In *Proceedings of the second ACM SIGSPATIAL international workshop on crowdsourced and volunteered geographic information*. 31–38.
[52] Song Gao, Linna Li, Wenwen Li, Krzysztof Janowicz, and Yue Zhang. 2017. Constructing gazetteers from volunteered big geo-data based on Hadoop. *Computers, Environment and Urban Systems* 61 (2017), 172–186.
[53] Song Gao, Hao Yu, Yong Gao, and Yinle Sun. 2010. A design of RESTful style digital gazetteer service in cloud computing environment. In *2010 18th International Conference on Geoinformatics*. IEEE, 1–6.
[54] Judith Gelernter, Gautam Ganesh, Hamsini Krishnakumar, and Wei Zhang. 2013. Automatic gazetteer enrichment with user-geocoded data. In *Proceedings of the Second ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information*. 87–94.
[55] Hugh H Genoways and Suzanne B McLaren. 2003. Results of the Alcoa Foundation-Suriname Expeditions. XIII. Annotated gazetteer of mammal collecting sites in Suriname. *ANNALS-CARNEGIE MUSEUM PITTSBURGH* 72, 4 (2003), 223–240.
[56] Daniel W Goldberg, John P Wilson, and Craig A Knoblock. 2009. Extracting geographic features from the internet to automatically build detailed regional gazetteers. *International Journal of Geographical Information Science* 23, 1 (2009), 93–128.
[57] Michael F Goodchild. 2007. Citizens as sensors: the world of volunteered geography. *GeoJournal* 69 (2007), 211–221.
[58] Michael F. Goodchild. 2016. Gazetteers Present: Spatial Science and Volunteered Geographical Information. In *Placing Names: Enriching and Integrating Gazetteers*. Indiana University Press, 26–34. http://www.jstor.org/stable/j.ctt2005zq7.7
[59] Mark Graham and Stefano De Sabbata. 2015. Mapping information wealth and poverty: the geography of gazetteers. *Environment and Planning A* 47, 6 (2015), 1254–1264.
[60] Lenore A Grenoble, Hilary McMahan, and Alliaq Kleist Petrussen. 2019. An ontology of landscape and seascape in Greenland: The linguistic encoding of land in Kalaallisut. *International Journal of American Linguistics* 85, 1 (2019), 1–43.
[61] Karl Grossner and Ruth Mostern. 2021. Linked places in world historical gazetteer. In *Proceedings of the 5th ACM SIGSPATIAL International Workshop on Geospatial Humanities*. 40–43.
[62] Thomas R Gruber. 1995. Toward principles for the design of ontologies used for knowledge sharing? *International journal of human-computer studies* 43, 5-6 (1995), 907–928.
[63] Ehsan Hamzei, Stephan Winter, and Martin Tomko and. 2020. Place facets: a systematic literature review. *Spatial Cognition & Computation* 20, 1 (2020), 33–81. https://doi.org/10.1080/13875868.2019.1688332
[64] Shoichiro Hara. 2017. Digital gazetteer as a knowledgebase for open data science. In *2017 Pacific Neighborhood Consortium Annual Conference and Joint Meetings (PNC)*. IEEE, 69–75.
[65] Shoichiro Hara and Tatsuki Sekino. 2018. Digital Gazetteer as a Knowledgebase for Open Data Science (2 nd Report). In *2018 Pacific Neighborhood Consortium Annual Conference and Joint Meetings (PNC)*. IEEE, 1–6.
[66] JT Hastings. 2008. Automated conflation of digital gazetteer data. *International Journal of Geographical Information Science* 22, 10 (2008), 1109–1127.
[67] L Hill. 2006. Gazetteers and Gazetteer Services. *Georeferencing: The geographic associations of information* (2006), 91–154.
[68] Linda L Hill. 2000. Core elements of digital gazetteers: placenames, categories, and footprints. In *Research and Advanced Technology for Digital Libraries: 4th European Conference, ECDL 2000 Lisbon, Portugal, September 18–20, 2000 Proceedings 4*. Springer, 280–290.
[69] Linda L Hill and Qi Zheng. 1999. Indirect geospatial referencing through place names in the digital library: Alexandria digital library experience with the developing and implementing gazetteers: Analysis and preliminary evaluation of the classical digital library model. In *Proceedings of the Annual Meeting-American Society for Information Science*, Vol. 36. Citeseer, 57–69.
[70] Livia Hollenstein and Ross Purves. 2010. Exploring place through user-generated content: Using Flickr tags to describe city cores. *Journal of Spatial Information Science* 1 (2010), 21–48.
[71] Ryan Horne. 2020. Beyond lists: digital gazetteers and digital history. *The Historian* 82, 1 (2020), 37–50.
[72] Yingjie Hu, Huina Mao, and Grant McKenzie. 2019. A natural language processing and geospatial clustering framework for harvesting local place names from geotagged housing advertisements. *International Journal of Geographical*

Information Science 33, 4 (2019), 714–738.

[73] I-Mei Hung, Yu-Ji Li, Pi-Ling Pai, and Hsiung-Ming Liao. 2008. The construction of the chinese gazetteer information system: the integration application of authority control, gazetteer and GIS. In *Archiving Conference*, Vol. 5. Society of Imaging Science and Technology, 246–256.

[74] Suela Isaj, Esteban Zimányi, and Torben Bach Pedersen. 2019. Multi-source spatial entity linkage. In *Proceedings of the 16th International Symposium on Spatial and Temporal Databases*. 1–10.

[75] Krzysztof Janowicz and Carsten Keßler. 2008. The role of ontology in improving gazetteer interaction. *International Journal of Geographical Information Science* 22, 10 (2008), 1129–1157.

[76] Christopher B Jones, Harith Alani, and Douglas Tudhope. 2001. Geographical information retrieval with ontologies of place. In *Spatial Information Theory: Foundations of Geographic Information Science International Conference, COSIT 2001 Morro Bay, CA, USA, September 19–23, 2001 Proceedings 5*. Springer, 322–335.

[77] Christopher B Jones, Ross S Purves, Paul D Clough, and Hideo Joho. 2008. Modelling vague places with knowledge from the Web. *International Journal of Geographical Information Science* 22, 10 (2008), 1045–1065.

[78] Susan L Kelley. 2006. Resolving place names in Amdo and Kham: A gazetteer for the Hengduan Mountains region of Southwest China. *Journal of Systematics and Evolution* 44, 6 (2006), 721.

[79] Helen Kerfoot. 2016. Gazetteers Global: United Nations Geographical Name Standardization. In *Placing Names: Enriching and Integrating Gazetteers*. Indiana University Press, 35–50. http://www.jstor.org/stable/j.ctt2005zq7.8

[80] Carsten Keßler, Krzysztof Janowicz, and Mohamed Bishr. 2009. An agenda for the next generation gazetteer: Geographic information contribution and retrieval. In *Proceedings of the 17th ACM SIGSPATIAL international conference on advances in Geographic Information Systems*. 91–100.

[81] Carsten Keßler, Patrick Maué, Jan Torben Heuer, and Thomas Bartoschek. 2009. Bottom-up gazetteers: Learning from the implicit semantics of geotags. In *GeoSpatial Semantics: Third International Conference, GeoS 2009, Mexico City, Mexico, December 3-4, 2009. Proceedings 3*. Springer, 83–102.

[82] Johnathan P Kirk and Gordon A Cromley. 2018. Assimilating weather data into a digital event gazetteer of airborne parachute operations during the French Indochina War. *Weather, climate, and society* 10, 1 (2018), 19–34.

[83] George Lamprianidis, Dimitrios Skoutas, George Papatheodorou, and Dieter Pfoser. 2014. Extraction, integration and analysis of crowdsourced points of interest from multiple web sources. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information*. 16–23.

[84] Ora Lassila, J Hendler, and T Berners-Lee. 2001. The semantic web. *Scientific American* 284, 5 (2001), 34–43.

[85] Robert Laurini. 2015. Geographic ontologies, gazetteers and multilingualism. *Future Internet* 7, 1 (2015), 1–23.

[86] Huali Li, Jun Liu, and Xiran Zhou. 2018. Intelligent map reader: A framework for topographic map understanding with deep learning and gazetteer. *IEEE Access* 6 (2018), 25363–25376.

[87] Linna Li and Michael F Goodchild. 2012. Constructing places from spatial footprints. In *Proceedings of the 1st ACM SIGSPATIAL international workshop on crowdsourced and volunteered geographic information*. 15–21.

[88] Michael D Lieberman, Hanan Samet, and Jagan Sankaranarayanan. 2010. Geotagging with local lexicons to build indexes for textually-specified spatial data. In *2010 IEEE 26th international conference on data engineering (ICDE 2010)*. IEEE, 201–212.

[89] Nung-yao Lin, Shih-pei Chen, Sean Wang, and Calvin Yeh. 2020. Displaying spatial epistemologies on web GIS: using visual materials from the Chinese local gazetteers as an example. *International Journal of Humanities and Arts Computing* 14, 1-2 (2020), 81–97.

[90] Jeffrey Liu and Ziling Wan. 2022. The Making of a Sacred Landscape: Visualizing Hangzhou Buddhist Culture via Geoparsing a Local Gazetteer the Xianchun Lin'an zhi. *Religions* 13, 8 (2022), 711.

[91] Sophia B Liu, Leysia Palen, Jeannette Sutton, Amanda L Hughes, Sarah Vieweg, et al. 2008. In search of the bigger picture: The emergent role of on-line photo sharing in times of disaster. In *Proceedings of the information systems for crisis response and management conference (ISCRAM)*. Citeseer, 4–7.

[92] Yu Liu, Runqiang Li, Kaichen Chen, Yihong Yuan, Lingli Huang, and Hao Yu. 2009. KIDGS: A geographical knowledge-informed digital gazetteer service. In *2009 17th International Conference on Geoinformatics*. IEEE, 1–6.

[93] Britt Lonneville, Harm Delva, Marc Portier, Laurian Van Maldeghem, Lennert Schepers, Dias Bakeev, Bart Vanhoorne, Lennert Tyberghein, and Pieter Colpaert. 2021. Publishing the Marine Regions Gazetteer as a Linked Data Event Stream.. In *JOWO*.

[94] Peter Lucas, Magesh Balasubramanya, Dominic Widdows, and Michael Higgins. 2006. The Information Commons Gazetteer.. In *LREC*. 1746–1751.

[95] Ivre Marjorie Machado, Rafael Odon de Alencar, Roberto de Oliveira Campos Junior, and Clodoveu A Davis Jr. 2010. An Ontological Gazetter for Geographic Information Retrieval.. In *GeoInfo*. 21–32.

[96] Hugo Manguinhas, Bruno Martins, and José Borbinha. 2008. A geo-temporal web gazetteer integrating data from multiple sources. In *2008 Third international conference on digital information management*. IEEE, 146–153.

[97] David M Mark and Andrew G Turk. 2003. Ethnophysiography. In *Workshop on Spatial and Geographic Ontologies*.

[98] Bruno Martins. 2011. A supervised machine learning approach for duplicate detection over gazetteer records. In *International Conference on GeoSpatial Sematics*. Springer, 34–51.
[99] Bruno Martins, Helena Galhardas, and Nelson Gonçalves. 2012. Using Random Forest classifiers to detect duplicate gazetteer records. In *7th Iberian Conference on Information Systems and Technologies (CISTI 2012)*. IEEE, 1–4.
[100] Patrick Maué. 2007. Reputation as tool to ensure validity of VGI. In *Workshop on volunteered geographic information*.
[101] Katherine McDonough and Matje van de Camp. 2017. Mapping the encyclopédie: working towards an early modern digital gazetteer. In *Proceedings of the 1st ACM SIGSPATIAL Workshop on Geospatial Humanities*. 16–22.
[102] Kevin McDougall. 2009. The potential of citizen volunteered spatial information for building SDI. In *GSDI 11 world conference: spatial data infrastructure convergence: building SDI bridges to address global challenges*.
[103] Scott R McEathron, Patrick McGlamery, Dong-Guk Shin, Ben Smith, and Yuan Su. 2001. Naming the Landscape: Building the Connecticut Digital Gazetteer. (2001).
[104] Deborah L McGuinness, Frank Van Harmelen, et al. 2004. OWL web ontology language overview. *W3C recommendation* 10, 10 (2004), 2004.
[105] Fernando Melo and Bruno Martins. 2017. Automated geocoding of textual documents: A survey of current approaches. *Transactions in GIS* 21, 1 (2017), 3–38.
[106] Ruth Mostern. 2008. Historical gazetteers: An experiential perspective, with examples from Chinese history. *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 41, 1 (2008), 39–46.
[107] Ruth Mostern and Humphrey Southall. 2016. Gazetteers Past: Placing Names from Antiquity to the Internet. In *Placing Names: Enriching and Integrating Gazetteers*. Indiana University Press, 15–25. http://www.jstor.org/stable/j.ctt2005zq7.6
[108] Tiago HVM Moura and Clodoveu A Davis Jr. 2014. Integration of linked data sources for gazetteer expansion. In *Proceedings of the 8th Workshop on Geographic Information Retrieval*. 1–8.
[109] Yoshikatsu Nagata. 2019. Community Level Old Place Names in the Northeast of Thailand for a Historical Digital Gazetteer. In *2019 Pacific Neighborhood Consortium Annual Conference and Joint Meetings (PNC)*. IEEE, 1–6.
[110] Shawn Newsam and Yi Yang. 2008. Integrating gazetteers and remote sensed imagery. In *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems*. 1–10.
[111] Giang Nguyen, Štefan Dlugolinskỳ, Michal Laclavík, Martin Šeleng, and Viet Tran. 2014. Next improvement towards linear named entity recognition using character gazetteers. In *Advanced Computational Methods for Knowledge Engineering: Proceedings of the 2nd International Conference on Computer Science, Applied Mathematics and Applications (ICCSAMA 2014)*. Springer, 255–265.
[112] L Obrst. 2010. *The Ontology Spectrum*. Springer Netherlands.
[113] Maxwell Guimarães de Oliveira, Cláudio EC Campelo, Cláudio de Souza Baptista, and Michela Bertolotto. 2016. Gazetteer enrichment for addressing urban areas: A case study. *Journal of Location Based Services* 10, 2 (2016), 142–159.
[114] Pi-Ling Pai and I-Chun Fan. 2016. Gazetteer GIS and the Study of Taiwan Local Society and Its Transition. In *Placing Names: Enriching and Integrating Gazetteers*. Indiana University Press, 217–230. http://www.jstor.org/stable/j.ctt2005zq7.20
[115] Xiaobo Peng, Rongguo Chen, Changxiu Cheng, and Xun Yan. 2010. A folksonomy-ontology-based digital gazetteer service. In *2010 18th International Conference on Geoinformatics*. IEEE, 1–6.
[116] Du Ping and Liu Yong. 2009. Building place name ontology to assist in geographic information retrieval. In *2009 International Forum on Computer Science-Technology and Applications*, Vol. 1. IEEE, 306–309.
[117] Mark Polczynski and Michael Polczynski. 2022. Lessons learned from using historical maps to create a digital gazetteer of historical places. *International Journal of Cartography* 8, 3 (2022), 326–342.
[118] Adrian Popescu, Gregory Grefenstette, and Houda Bouamor. 2009. Mining a multilingual geographical gazetteer from the web. In *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, Vol. 1. IEEE, 58–65.
[119] Adrian Popescu, Gregory Grefenstette, and Pierre Alain Moëllic. 2008. Gazetiki: automatic creation of a geographical gazetteer. In *Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries*. 85–93.
[120] S Pradeepa and KR Manjula. 2016. Construction of gazetteers from geo big data using machine learning technique on Hadoop. In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*. IEEE, 1619–1622.
[121] Eric Prud'hommeaux and Andy Seaborne. 2008. SPARQL Query Language for RDF. https://www.w3.org/TR/rdf-sparql-query/
[122] Ross S Purves, Paul Clough, Christopher B Jones, Mark H Hall, Vanessa Murdock, et al. 2018. Geographic information retrieval: Progress and challenges in spatial search of text. *Foundations and Trends® in Information Retrieval* 12, 2-3 (2018), 164–318.

[123] Ross S. Purves, Stephan Winter, and Werner Kuhn. 2019. Places in Information Science. *J. Assoc. Inf. Sci. Technol.* 70, 11 (Oct. 2019), 1173–1182. https://doi.org/10.1002/asi.24194

[124] Tye Rattenbury, Nathaniel Good, and Mor Naaman. 2007. Towards automatic extraction of event and place semantics from flickr tags. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. 103–110.

[125] Blake Regalia, Krzysztof Janowicz, Gengchen Mai, Dalia Varanka, and E Lynn Usery. 2018. GNIS-LD: Serving and visualizing the geographic names information system gazetteer as linked Data. In *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15*. Springer, 528–540.

[126] James Reid. 2003. geoXwalk–A Gazetteer Server and Service for UK Academia. In *International Conference on Theory and Practice of Digital Libraries*. Springer, 387–392.

[127] Matthew T Rice, Ahmad O Aburizaiza, R Daniel Jacobson, Brandon M Shore, and Fabiana I Paez. 2012. Supporting Accessibility for Blind and Vision-impaired People With a Localized Gazetteer and Open Source Geotechnology. *Transactions in GIS* 16, 2 (2012), 177–190.

[128] Wolf-Fritz Riekert. 2002. Automated Retrieval of Information in the Internet by Using Thesauri and Gazetteers as Knowledge Sources. *Journal of Universal Computer Science* 8, 6 (2002), 581–590.

[129] Vivek Sehgal, Lise Getoor, and Peter D Viechnicki. 2006. Entity resolution in geospatial data integration. In *Proceedings of the 14th annual ACM international symposium on Advances in geographic information systems*. 83–90.

[130] Hansi Senaratne, Amin Mobasheri, Ahmed Loai Ali, Cristina Capineri, and Mordechai Haklay. 2017. A review of volunteered geographic information quality assessment methods. *International Journal of Geographical Information Science* 31, 1 (2017), 139–167.

[131] Khaled Shaalan and Hafsa Raza. 2008. Arabic named entity recognition from diverse text types. In *Advances in Natural Language Processing: 6th International Conference, GoTAL 2008 Gothenburg, Sweden, August 25-27, 2008 Proceedings*. Springer, 440–451.

[132] Pavel Shvaiko and Jérôme Euzenat. 2005. A survey of schema-based matching approaches. In *Journal on data semantics IV*. Springer, 146–171.

[133] Raj Singh. 2016. International Standards for Gazetteer Data Structures. In *Placing Names: Enriching and Integrating Gazetteers*. Indiana University Press, 67–79. http://www.jstor.org/stable/j.ctt2005zq7.10

[134] Sanket Kumar Singh and Davood Rafiei. 2018. Strategies for geographical scoping and improving a gazetteer. In *Proceedings of the 2018 World Wide Web Conference*. 1663–1672.

[135] Philip D Smart, Christopher B Jones, and Florian A Twaroch. 2010. Multi-source toponym data integration and mediation for a meta-gazetteer service. In *Geographic Information Science: 6th International Conference, GIScience 2010, Zurich, Switzerland, September 14-17, 2010. Proceedings 6*. Springer, 234–248.

[136] Humphrey Southall, Alexander Von Luenen, and Paula Aucott. 2009. On the organisation of geographical knowledge: data models for gazetteers and historical GIS. In *2009 5th IEEE International Conference on E-Science Workshops*. IEEE, 162–166.

[137] Ligiane A Souza, Clodoveu A Davis, Karla AV Borges, Tiago M Delboni, and Alberto HF Laender. 2005. The role of gazetteers in geographic knowledge discovery on the web. In *Third Latin American Web Congress (LA-WEB'2005)*. IEEE, 9–pp.

[138] Evan A Sultanik and Clayton Fink. 2012. Rapid geotagging and disambiguation of social media text via an indexed gazetteer.. In *ISCRAM*.

[139] William Sunna and Isabel F Cruz. 2007. Structure-based methods to enhance geospatial ontology alignment. In *International Conference on GeoSpatial Sematics*. Springer, 82–97.

[140] Vlad Tanasescu, Philip D Smart, and Christopher B Jones. 2014. Reverse geocoding for photo captioning with a meta-gazetteer. In *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 509–512.

[141] Mark Thurstain-Goodwin and David Unwin. 2000. Defining and delineating the central areas of towns for statistical monitoring using continuous surface representations. *Transactions in GIS* 4, 4 (2000), 305–317.

[142] Klaus Tochtermann, Wolf-Fritz Riekert, Gerlinde Wiest, Jürgen Seggelke, and Birgit Klohaupt-Jahr. 1997. Using semantic, geographical, and temporal relationships to enhance search and retrieval in digital catalogs. In *Research and Advanced Technology for Digital Libraries: First European Conference, ECDL'97 Pisa, Italy, September 1–3, 1997 Proceedings 1*. Springer, 73–86.

[143] Yi-Fu Tuan. 1977. *Space and place: The perspective of experience.* U of Minnesota Press.

[144] Andrew Turner. 2006. *Introduction to neogeography.* " O'Reilly Media, Inc.".

[145] Florian A Twaroch, Paul Brindley, Paul D Clough, Christopher B Jones, Robert C Pasley, and Sue Mansbridge. 2019. Investigating behavioural and computational approaches for defining imprecise regions. *Spatial Cognition & Computation* 19, 2 (2019), 146–171.

[146] Florian A Twaroch and Christopher B Jones. 2010. A web platform for the evaluation of vernacular place names in automatically constructed gazetteers. In *Proceedings of the 6th Workshop on Geographic Information Retrieval*. 1–2.
[147] Florian A Twaroch, Christopher B Jones, and Alia I Abdelmoty. 2008. Acquisition of a vernacular gazetteer from web sources. In *Proceedings of the first international workshop on Location and the web*. 61–64.
[148] Florian A Twaroch, Philip D Smart, and Christopher B Jones. 2008. Mining the web to detect place names. In *Proceedings of the 5th Workshop on Geographic Information Retrieval*. 43–44.
[149] Mark Van Assem, Maarten R Menken, Guus Schreiber, Jan Wielemaker, and Bob Wielinga. 2004. A method for converting thesauri to RDF/OWL. In *International Semantic Web Conference*. Springer, 17–31.
[150] M van Van Exel, Eduardo Dias, and Steven Fruijtier. 2010. The impact of crowdsourcing on spatial data quality indicators. In *Proceedings of the GIScience 2010 Doctoral Colloquium, Zurich, Switzerland*. 14–17.
[151] Øyvind Vestavik and Ingeborg T Sølvberg. 2007. Merging local and global gazetteers. In *Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers: 10th International Conference on Asian Digital Libraries, ICADL 2007, Hanoi, Vietnam, December 10-13, 2007. Proceedings 10*. Springer, 495–496.
[152] Valeria Vitale, Pau de Soto, Rainer Simon, Elton Barker, Leif Isaksen, and Rebecca Kahn. 2021. Pelagios–connecting histories of place. Part I: Methods and tools. *International Journal of Humanities and Arts Computing* 15, 1-2 (2021), 5–32.
[153] Jun Wang and Ning Ge. 2006. Automatic feature thesaurus enrichment: extracting generic terms from digital gazetteer. In *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*. 326–333.
[154] Mathew Weaver, Lois Delcambre, Leonard Shapiro, Jason Brewster, Afrem Gutema, and Timothy Tolle. 2003. A digital geolibrary: Integrating keywords and place names. In *Research and Advanced Technology for Digital Libraries: 7th European Conference, ECDL 2003 Trondheim, Norway, August 17-22, 2003 Proceedings 7*. Springer, 422–433.
[155] Mats Widgren. 2018. Mapping global agricultural history: A map and gazetteer for sub-Saharan Africa, c. 1800 AD. *Plants and people in the African past: Progress in African archaeobotany* (2018), 303–327.
[156] John Wilson. 1882. *The Gazetteer of Scotland*. Edinburgh W. A.K. Johnston.
[157] John P Wilson, Christine S Lam, and Deborah A Holmes-Wong. 2004. A new method for the specification of geographic footprints in digital gazetteers. *Cartography and Geographic Information Science* 31, 4 (2004), 195–207.
[158] Liping Yang, Guangfa Lin, Ailing Chen, Youfei Chen, and Xiaohuan Wen. 2010. A spatio-temporal data model for administrative division place names: a case study of Xiamen. In *Sixth International Symposium on Digital Earth: Models, Algorithms, and Virtual Reality*, Vol. 7840. SPIE, 73–82.
[159] Du Yongtao. 2012. Literati and spatial order: A preliminary study of comprehensive gazetteers in the late Ming. *Ming Studies* 2012, 66 (2012), 16–43.
[160] Nagata Yoshikatsu. 2017. Geographic names on old maps of early 20th century toward a spatio-temporal gazetteer: A study on their accuracy in Northeast Thailand. In *2017 Pacific Neighborhood Consortium Annual Conference and Joint Meetings (PNC)*. IEEE, 98–103.
[161] Masaharu Yoshioka and Takahiro Fujiwara. 2013. Construction of a Japanese gazetteers for Japanese local toponym disambiguation. In *Proceedings of the 7th Workshop on Geographic Information Retrieval*. 57–63.
[162] Hee Cheon Yun, Joon Kyu Park, and Jong Sin Lee. 2013. Efficient Registration Plan of Place Names for Reinforcement of Active Region in Antarctica. *Journal of the Korean Society of Surveying, Geodesy, Photogrammetry and Cartography* 31, 6_2 (2013), 549–557.
[163] Kiran Zahra. 2017. Natural Disaster Database Design and Development for Himalaya Using Social Media.. In *AGILE PhD School*.
[164] Haihui Zhang, Yuanziyi Zhang, Qi Xin, and Fanghui Xiao. 2020. Contemporary Chinese Village Gazetteer Data Project: From Books to Data. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*. 559–560.
[165] Xueying Zhang and Shaonan Zhu. 2010. Contextual spatial relations based spatial modelling of vague place names. In *2010 Second IITA International Conference on Geoscience and Remote Sensing*, Vol. 1. IEEE, 23–26.
[166] Yu Zheng, Xixuan Fen, Xing Xie, Shuang Peng, and James Fu. 2010. Detecting nearly duplicated records in location datasets. In *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*. 137–143.
[167] Changqing Zhou, Dan Frankowski, Pamela Ludford, Shashi Shekhar, and Loren Terveen. 2004. Discovering personal gazetteers: an interactive clustering approach. In *Proceedings of the 12th annual ACM international workshop on Geographic information systems*. 266–273.

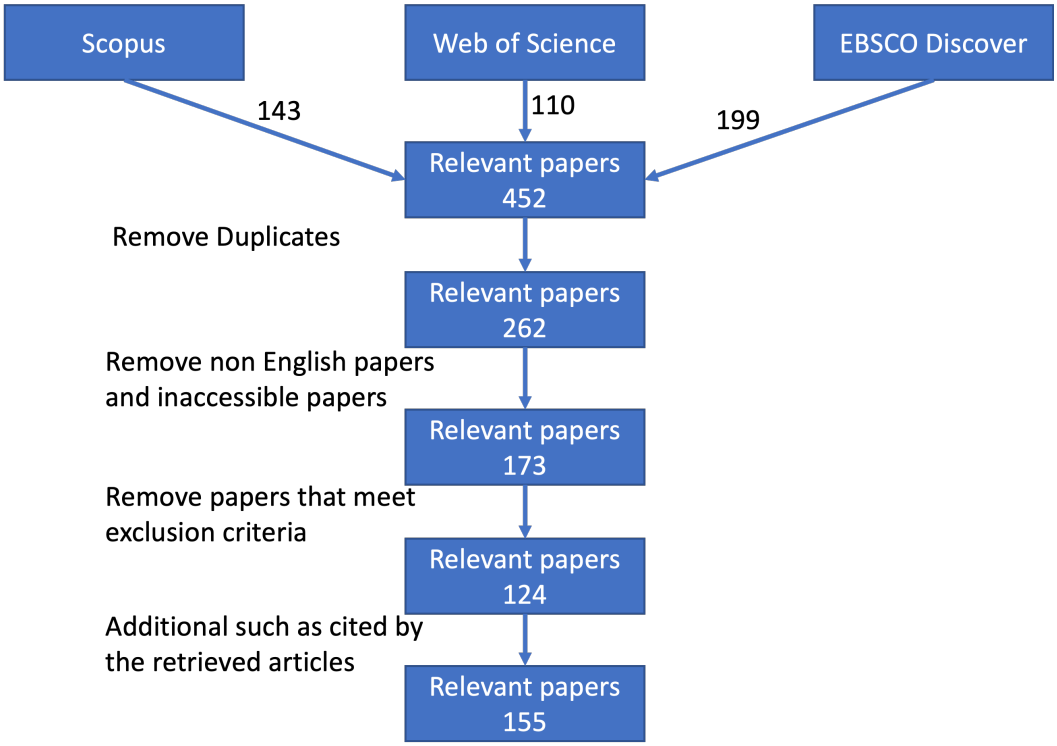

Fig. 7. Breakdown of the papers retrieved from the two search queries.

Table 5. Summary and comparison of papers that discuss duplicate identification and the methods used. Each number in the first column refers to a citation.

Paper	Duplicate Identification	Source/ Sources	Components used for matching			Notes
			Feature Name	Geographic Footprint	Feature Type	
[69]	Inter-gazetteer	USGS Geographic Name Service and NIMA GeoNames server	-	-	-	Uses GeoNames server for places outside the US and USGS Geographic Names Service for places within the US therefore avoiding duplicates
[94]	Inter-gazetteer	GeoNames server and GNIS	-	-	-	Uses GNS for places outside of the US and GNIS for places within the US therefore avoiding duplicates
[129]	Inter and Intra gazetteer	United Kingdom Permanent Committee on Geographic Names and Geographic Names Database of the United States Board on Geographic Names	Edit distance, Jaccard and Jaro-Winkler similarity measures	Inverse geographic distance between coordinate pairs	Type similarity based on co-occurrence in the manually annotated dataset they have prepared	Features are fed into logistic regression, voted perceptron (Neural Network) and support vector machine models. Logistic regression achieves best results
[151]	Inter-gazetteer	ADL and Sentralt Stedsnavnregister (SSR) (Official Norwegian place registry)	Number of shared place names	Proximity and whether contained by same larger feature	Hand crafted mapping between two thesauri. Similarity measured as a graph distance over two vocabularies.	
[66]	Inter and Intra gazetteer	Geographic Data Technology (GDT, now TeleAtlas), GNIS, ESRI 'Data and Maps CD', 'Transportation 2.0' data suite, Lake Tahoe Data Clearing house	A stemming string similarity algorithm	Collocation used as a threshold for candidate selection. Area of overlap of footprints used for matching	The steps to the common ancestor in the feature type hierarchy	Proposes methods for duplicate identification. Demonstrates with the mentioned sources. Matching of different thesauri is out of scope

[96]	Inter-gazetteer	Geonames, GeoNetPT, Google Earth Community, Estonian National Library, ECAI time period directory, Wikipedia	Manually tuned threshold for Jaro-Winkler token similarity	Distance below a manually tuned threshold	Introduce a novel ontology which features from all sources are manually mapped to	-
[33]	Inter and intra gazetteer	National Geographical Names Database and geographical names information system issued by Chinese Ministry of Civil affairs	Threshold for token similarity	Distance below a threshold containment	Manually maps sources to a reference ontology and obtains a semantic similarity	Uses a weighted sum of the three types of similarities similar to Hastings, 2008.
[166]	Intra-gazetteer	Bing Maps	Edit distances and inverse document frequency of similar and dissimilar strings between names	Uses address instead of a footprint. Builds an address hierarchy and calculates distance to coparent.	The levels to reach the coparent for the two features in the category hierarchy	Uses a decision tree classifier for matching
[135]	Inter-gazetteer	GeoNames, OSM, Yahoo WOE, Wikipedia, OS Point X, OS 50k, OS MasterMap	Levenshtein edit distance, Soundex, Text normalisation	Distance threshold applied to distances between geometry objects of points and polygons		A threshold is applied to a weighted combination of Levenshtein and Soundex distances before applying a spatial distance threshold.
[98]	Intra-gazetteer	ADL and other unspecified sources	Variety of string similarity features including Levenshtein, Jaro-Winkler and Monge-Elkan distances	Variety of distance (minimum distance between footprints, distance between centroids, etc), overlap of features, overlap relative to feature size, etc	Variety of features like Jaccard coefficient, Dice coefficient, etc of place types associated with the place, up-steps to common ancestor	Features fed into an SVM and alternating decision tree classifier. Author has also considered semantic relations and temporal features. Method proposed however is only applicable to features from the same feature type thesauri
[54]	Inter-gazetteer	GeoNames, Wikimapia, OpenStreetMap	Weighted N-gram similarity, edit distance, soundex matching scores	Normalized geographic distance	-	Uses two SVMs to learn weights for the n-gram queries for place names and to rank the candidates. Gives a fuzzy result.

[108]	Inter-gazetteer	GeoNames, DBPedia	Only exact name matches are considered	containedBy predicate is used to check common places that contain the two candidates	-	Exploit semantic tags and Wikipedia links. The sameAs predicate or the sharing of a common wikipedia link is considered to be an exact match between the two linked data entries.
[74]	Inter and Intra POI	Google Places, Foursquare, Yelp, and Krak	Levenshtein string similarity	Filtering with an adapted quadtree	Wu-Palmer similarity applied to pairs of attributes from different entities	Non-supervised pareto optimisation applied to text and semantic similarity measures in combination with determining threshold for skyline to distinguish positive / negative matches.
[2]	Inter-Gazetteer	Geonames, SwissNames3D	Levenshtein-Damerau distance, the Jaro similarity, and the Jaro-Winkler similarity. Also uses Levenshtein distance on any alternate names	point-to-point distance between gazetteer records	one-hot encoding to encode feature types	Features fed into a random forest classifier to classify pairs of places. Also use other attributes like elevation and land cover. Also compares results with a rule based approach
[11]	Inter-Gazetteer	Yelp, Foursquare and OSM	BERT embedding of text	A distance embedding that is a function of Haversine distance	-	A blocking phase selects pairs that meet a string similarity and distance threshold. Also uses a contextual embedding based on combining BERT embeddings of target entities with nearby entities. All embeddings combined to make predictions