Design and implementation of a *Gazetteer*

André Soares - 55351
Universidade Técnica de Lisboa
Instituto Superior Técnico
Lisboa, Portugal
andre.soares@ist.utl.pt

ABSTRACT

The objective of this work is the analysis and subsequent remodeling of a Gazetteer service from a previous project (DIGMAP) in order to adapt it to the needs of a European project (EuropeanaConnect). The goal is to produce a Gazetteer service for use by another service dedicated to scan text for geographic references - a Geoparser. The end result was a service that complies with current standards and practices defined for a service of its nature. First a detailed description of the concept of a Gazetteer and its history is presented along with any other relevant concepts for the purposes of understanding the performed work. This is followed by a survey of several related initiatives in the area of Gazetteers that were relevant in developing concepts, methods and technologies for that area. Then a study of the more pertinent issues currently in research for the purposes of Gazetteer development. Afterwards, we give a definition of the actual problem and requirements for the purposes of this work followed by a description of the planned architecture for the Gazetteer and the actual implementation. Next we test some components of the system, presenting statistics regarding the data stored in the gazetteer, to determine its quality and reliability. We finish with a description of end results and personal conclusions.

Keywords

Gazetteer; Geographic Information; Geographic Information System

1. INTRODUCTION

A gazetteer consists of a list of geographic names, together with their geographic locations and other descriptive information [12]. Nowadays, with the wide use of the Internet, there is an increasing search and need from both applications and users for this type of information, be it for driving directions, buying houses or planning foreign travel. This, coupled with the recent potential for users to create their own geographic information (e.g. Flickr[1], Wikimapia[2]), makes gazetteers becoming increasingly important as tools for providing geographic context on the Web.

The motivation for this work was to construct a new design, implementation and validation of a gazetteer initially developed for the DIGMAP project [4], based upon findings of new requirements and improvements to the actual state of the art in technology for gazetteers. The ensuing work was performed on the context of the Europeana Connect project[3] which was responsible for producing core components essential for the realization of the Europeana service[4], so as to become a truly interoperable, multilingual and user-oriented service.

One of the ways this is accomplished is by further enriching the content provided by Europeana by adding information about geographic locations featured in the digital objects. This is done by the use of two Geographic Information Systems (GIS): a geoparser and a gazetteer. Because Europeana processes metadata records from a large number of data providers and aggregators, the information present in those records can be enhanced through the use of the Geoparser service which extracts from the records geographic references found in unstructured text so these can be embodied in the records as new structured elements.

However, the Geoparser does not possess any form of geographic information regarding the references it detects and therefore, can not be used by itself to create new useful information for the records. This is where the Gazetteer steps in. Being a geographic information database, the Gazetteer can also serve as a stand-alone service for geographical information browsing and searching. Also, it provides methods for managing and adding new information into the gazetteer, serving as an aggregator for various geographical data providers and consolidating that information into less and more complete records, removing both duplicate and irrelevant data. With the help of these services, new information is added into Europeana records, providing more ways to search and organize the information contained therein.

The results gathered from this work show that the resulting Gazetteer System system has already established the main structure and components required for complying with Europeana's requirements and, as such can be used as production component in the Europeana service. The system is comprised by two parts: first there's the main ISO and OGC compliant service for geographic content discovery and then there is the addition of an "ecology" of added services designed to enhance the Gazetteer's functionality and capabilities much like the GIS Suite this Gazetteer service is a part of, is designed to enhance the search and content relationship capabilities of the Europeana service. These in-

[1] http://www.flickr.com
[2] http://wikimapia.org
[3] http://www.europeanaconnect.eu
[4] http://www.europeana.eu

clude a content Ingester for importing content from several data providers and a Duplicate Detection service, designed for quality control over the Gazetteer's content. The main contribution of the project is the addition of this services' ecology, which serves as a sort of "back-office" for the standard Gazetteer service, in order to improve its performance and reliability.

The remainder of this document is structured as follows. First, a brief look into the current state of the art in gazetteer development, along with some relevant initiatives. Next, a discussion of some of the more pertinent concerns when developing a gazetteer system followed by a description of the system developed and finally, a statistical classifying of data imported into the Gazetteer system and testing of the duplicate detection mechanism, the conclusions based upon the accomplished work and future work planning.

2. CONCEPT AND ROLE

Gazetteers are not exactly a recent concept, but their use in a digital perspective can be considered recent since there have yet to exist full specification and implementation standards. However, significant developments in projects during the last decade, followed by efforts from the ISO Technical Committee 211 (ISO) and the Open Geospatial Consortium (OGC), have led us to a point where we have a general agreement of their basic structure, attributes and basic protocols. Using some ISO standards combined with OGC specifications as a starting point, coupled with remarks from other papers, a description of the concept of a gazetteer will be presented. An important aspect of gazetteers are the entities they store, features [1]:

1. a **feature** is an abstraction of a real world phenomenon. It is a geographic feature if it is associated with a location relative to the Earth, in other words, if it contains information concerning phenomena implicitly or explicitly associated with a location relative to the Earth.

More concretely, they are distinct physical elements or objects (buildings, rivers, mountains) for which geographic positions and boundaries can be established [8]. Geographic features contain spatial references that relate them to positions on Earth and can be one of two types:

- coordinates
- geographic identifiers (sometimes referred to as indirect spatial references)

A geographic identifier can be in the form of a name (e.g. country name) or a label (e.g. postal code). When a geographic identifier is used as a spatial reference, it uniquely identifies a location [2]. In this context:

2. a **location** is a position on Earth which has an identity, be it as a city (Paris), country (France), monument (Eiffel Tower) or also places with vague positions and boundaries (e.g. Southern France).

This location is also a geographic feature and, as such, can be used to reference other features. So we can consider geographic features to be location instances. Usually, the geographic identifier shares a relationship with a location of containment within (e.g. Paris is contained in France) but there can be more complex relationships between them such as adjacency.

Spatial references can be organized into systems for identifying position in the real world. These are called spatial referencing systems and differ depending on the type of spatial reference used. For the purposes of the gazetteer definition, we will mainly consider the referencing system using geographic identifiers. This system consists of a related set of location types with their geographic identifiers. The relation between location types may form a hierarchy. Examples of these systems can be a list of countries, where the location type is country and the geographic identifiers can be the country name or code; another example is a set of addresses in a town, where the type in this case is the property and its geographic identifier is its respective address. With these concepts defined, a definition of a gazetteer [2] can now be introduced :

3. a **gazetteer** is a directory of geographic features, which are instances of one or more classes of location types, describing locations along with their position and additional information. The positional information (spatial reference) can be either coordinates or descriptive. If it contains a coordinate reference, this will allow relating from a spatial referencing system using geographic identifiers to a coordinate reference system. If it contains a descriptive reference, this will be a spatial reference using a different referencing system with geographic identifiers (possibly one where the location type used is related and lower in hierarchy than the one referring to the feature).

There may be several different gazetteers referring to the same location type, with the location instances identified in different ways. Conversely, a single gazetteer may include several location types. A gazetteer can be used both as an independent service simply to perform queries of locations ("Where is Lisbon?") or as a embedded resource aid for other information systems where the information they store is not directly related to specific locations. In those cases, the gazetteer first provides with a translation of a location to a set of coordinates or a bounding box (two points of a box that includes the spatial extent of the object), which can then be used by the system to obtain the desired information regarding a particular area. This sort of use for gazetteer data is commonly referred to as *indirect spatial referencing* [14]. In the latter case, they are becoming critical components in several types of information systems like Web-based mapping services, navigation services, geoparsers or spatial search engines. Also, they have been used as essential tools in problems of disambiguation of geographic names for problems of natural language parsing or collecting them for georeferencing contents [30, 33, 24].

3. RELATED INITIATIVES

In this section we will describe some of the most relevant projects that were related with gazetteers and have produced guidelines, technical results, or have been serving as main historical references in the context of gazetteer development.

3.1 ADL-G - Alexandria Digital Library Gazetteer

The ADL Project at the University of California, Santa Barbara was founded in 1994, under the support of the National Science Foundation (NSF)/Defense Advanced Research Projects Agency (DARPA)/National Aeronautics and Space Administration (NASA) Research in Digital Libraries Initiative. Developed by a consortium of researchers, developers, and educators from various sectors, the project's objective was to construct a distributed georeferenced digital library - *geolibrary*. A *geolibrary* is an organized collection of digital objects of which one of their primary attributes is their spatial location, in other words, their footprints [14].

The project required a gazetteer in order to provide their spatial query functions so as to georeference the objects in the library. This was accomplished with the merging of databases acquired from two U.S. federal agencies - U.S. National Imagery and Mapping Agency (NIMA) with their *Geographic Names Processing System* (GNPS) and the U.S. Geological Survey with their *Geographic Names Information System* (GNIS). During the process of merging and importing the data, the developers identified the following issues:

- the need for a standard conceptual schema for gazetteer information, in order to facilitate the creating and sharing of data between sources, making them more interoperable.
- Also, it became necessary to create a standard type schema, so as to provide a rich and unique one for gazetteers to be able to make their information more identifiable and easy to categorize and also <zso there could be mappings between various type schemas [12].

The solution for these issues was the development of the ADL Content Standard (GCS) and the Feature Type Thesaurus (FTT) respectively.

A *content standard* defines a common set of terms and definitions for the purpose of documenting data. It establishes names of entities and grouping of entities (collectives), their definitions and value restrictions. Its purpose is to enable data sharing and distributed access through a common representation of data about data. Therefore content standards guide the development of *metadata*. The GCS was built following a model of *metadata* because it enabled the possibility of contribution to place definitions from multiple sources and allowed the aggregation of information from multiple gazetteers that shared the content standard [14].

The purpose of the FTT was to try and establish a common link between the various typing schemas by adding, to the greatest extent possible, all of the vocabulary of the other schemas either as preferred terms, based on their average use in reference sources or dictionaries, or as alternate names pointing to the related preferred terms. Thus it is possible to have consistent description of types of places and features across several gazetteers. With that in mind, terms from several sources were collected and evaluated, including feature categories used by NIMA, categorization terms from the *Getty Thesaurus of Geographic Names*, *Getty Art and Architecture Thesaurus*, definitions from several dictionaries, among others. With the results, a hierarchy was constructed, with a small set of top categories:

- Administrative Areas
- Hydrographic Features
- Land Parcels
- Manmade Features
- Physiographic Features
- Regions

and the preferred terms were added as narrower terms of these categories, that is, they were added as members of a related top category and so forth.

Another relevant aspect of the project was the development of a protocol for accessing gazetteer services. The ADL Gazetteer Protocol[5] was designed to encourage system interoperability and as such provided low-level services to be simple enough that they can be implemented by all gazetteers, yet powerful enough to be useful to clients for their own purposes and for combining into higher-level services. The services they provide in concrete are three:

- `get-capabilities`. This service returned a description of the overall capabilities of the gazetteer (services and query types it supports, etc);
- `query`. This service returned reports of gazetteer entries selected by a query. This query was expressed in a gazetteer query language defined for the protocol;
- `download`. This service returns reports of all gazetteer entries;

The reports may be of two types: normal or extended. The main difference between them is the amount of information they contain. The normal one contains only some elements of a gazetteer entry, namely its system identifier, names, footprints, relationships and type classifications. The extended report contains all information pertaining to a gazetteer entry. The ADL gazetteer was very significant as a project in its merit by developing innovative services such as the Gazetteer Protocol and content models (GCS and FTT) but also in the sense that it showed the importance of gazetteers as helpful and important tools for providing geographic context on digital libraries and thus helped push efforts in research and development.

3.2 Geonames

The Geonames project was founded by swiss software engineer Marc Wick and launched at the end of 2005. It serves as a free and open source geographical database, designed primarily for use by developers who want to integrate the project into their own web services and applications. It contains world-wide geographical data including names of places in various languages, elevation, population, and latitude / longitude coordinates. Currently it's perhaps the most widely accepted geographic resource and also one of the most used geographic databases if not the most used. It contains over 8 million geographical names and consists of 7 million unique features. The main challenge in running this service is dealing with a huge number of data providers and the absence of gazetteer standards. However, unlike in the ADL project, where they utilize a Content Standard and a Feature Type Thesauri, in Geonames they use an ontology [15] of geographic features:

[5] `http://www.alexandria.ucsb.edu/gazetteer/protocol/`

4. an **ontology** is an explicit specification of a concept used to achieve a shared and common understanding of a particular domain of interest.

Meaning you define concepts giving them attributes such that each concept is different from each other and that they can also be used to restrict subconcepts to comply to the same attributes (if Sea has an attribute defined as a body of water, then all subconcepts of Sea have to be bodies of water as well) [15]. In Geonames, this is done by means of the OWL Web Ontology Language. OWL is a knowledge representation language for producing ontologies based on RDF (Resource Description Framework). RDF was created by the W3C (World Wide Web Consortium) as a family of specifications for data models, being used for conceptual descriptions or information modeling. An example of the application of the ontology is the following URL for the french town Embrun (the associated content is shown in the Figure 1):

1. `http://sws.geonames.org/3020251/`
2. `http://sws.geonames.org/3020251/about.rdf`

The first URL returns an HTML page regarding the town, while the latter returns a RDF document with the description of all the information Geonames has about it. Also, the web server is configured to redirect requests from the concept URL to the document one so that web agents can see Geonames has information concerning the feature. Another aspect of the ontology is that all features in Geonames are interlinked in some form. Depending of the type, the following document links are available:

- children (countries for a continent, administrative subdivisions for a country)
- neighbors (neigbouring countries)
- adjacent features

Also in Figure 1, where we have the document pertaining to Embrun, an example of the third type of linking between features is present with the use of the attribute `nearbyfeatures`.

3.3 DIGMAP

The Discovering our Past World with Digitised Maps (DIGMAP) project [4] was a recent collaborative effort between universities (including the IST, which was the overall project coordinator) and European national libraries that provided with their collections and services.

It was designed to provide new solutions for *geolibraries*, especially those focused on historical resources (ancient maps and reference documents). This was done in the creation of a virtual library for geographic content, demonstrating innovative ideas on the development of services for recovery and visualization of historic resources with relevant geographic features, based on collective metadata retrieved from the national libraries and other relevant third-party metadata sources, describing those resources. These resources were recovered from various sources on the Internet including online digital libraries and describe physical or digital ancient maps or documents as well as any relevant related web site.

The results of this project were the portal[6] and its integrated services which help collect and organize all the data in comprehensive collections, browsing indexes and search functions by the use of specialized tools: a catalogue, a feature extractor/indexer from images (in this case, the maps) , a metadata repository, a gazetteer and a geo-parser.

The gazetteer is used by the geo-parser to help identify relevant geographic features on the ancient texts. It was developed based on the ADL project standard using the ADL-GP (ADL Gazetteer Protocol) communications protocol, though a content model was created using OWL, creating an ontology based on its description logics, to incorporate semantic meaning to its contents. The typing scheme was based on a combination of the FTT with the classification scheme for time periods from the ECAI Time Period Directory, so it could also incorporate temporal information on its contents. That is one of the main differences about the DIGMAP gazetteer, it tries to establish both spatial and temporal referencing for places and regions on Earth. Content was then retrieved from relevant sources such as Geonames, the GeoNetPT OWL ontology and time period information was gathered both from the ECAI Time Period Directory and from Wikipedia.

A feature in the gazetteer consists in a series of attributes, namely a unique identifier, list of associated names with a preferred one chosen (like in TGN), place type classification through use of several feature type classification ontologies (ADL, Geonames, etc.), spatial reference (GML points, polygons, etc.), a temporal reference, relationships with other features (uses those of the Geonames ontology) plus additional information and external references (Wikipedia, dbpedia, etc.).

The project is currently only a demonstrative service with its results available to any interested entity.

4. MAIN CONCERNS

4.1 Data Integration and Data Quality

Many of the current concerns in the quality and performance of gazetteer service tie in mainly to the types of data that they store: Place names, Types and Footprints. Regarded as the core of gazetteers [13], these components each have their own characteristics and complexities to take into account when using or studying a gazetteer.

Place names are considered the set of places or sections of places which have acquired authoritative names and refer to them during a certain time span. Generally, they do not identify an arbitrary place uniquely since a name can refer to multiple locations and likewise an arbitrary location can be referred to by various names by different people and in different ways through time.

Place types improve communicating about places, for example when providing directions, and also reasoning about them, because they serve as abstractions, defining a set of perceivable characteristics that we would encounter in a region in space associated with that place type. Categorizing places isn't always a straight forward process: though place

[6] `http://portal.digmap.eu`

```
<Feature rdf:about="http://sws.geonames.org/3020251/">
    <name xml:lang="fr">Embrun</name>
    <alternateName xml:lang="fr">Embrun, Hautes-Alpes</alternateName>
    <featureClass rdf:resource="http://www.geonames.org/ontology#P"/>
    <featureCode rdf:resource="http://www.geonames.org/ontology#P.PPL"/>
    <inCountry rdf:resource="http://www.geonames.org/countries/#FR"/>
    <population>7069</population>
    <postalCode>05200</postalCode>
    <wgs84_pos:alt>900</wgs84_pos:alt>
    <wgs84_pos:lat>44.5667</wgs84_pos:lat>
    <wgs84_pos:long>6.5000</wgs84_pos:long>
    <parentFeature rdf:resource="http://sws.geonames.org/3013738/"/>
    <nearbyFeatures rdf:resource="http://sws.geonames.org/3020251/nearby.rdf"/>
    <locationMap>http://www.geonames.org/3020251/embrun.html</locationMap>
    <wikipediaArticle rdf:resource="http://fr.wikipedia.org/wiki/Embrun_%28Hautes-Alpes%29"/>
    <wikipediaArticle rdf:resource="http://pl.wikipedia.org/wiki/Embrun"/>
    <wikipediaArticle rdf:resource="http://de.wikipedia.org/wiki/Embrun"/>
    <wikipediaArticle rdf:resource="http://en.wikipedia.org/wiki/Embrun%2C_Hautes-Alpes"/>
    <wikipediaArticle rdf:resource="http://it.wikipedia.org/wiki/Embrun"/>
    <wikipediaArticle rdf:resource="http://nl.wikipedia.org/wiki/Embrun"/>
    <owl:sameAs rdf:resource="http://rdf.insee.fr/geo/COM_05046"/>
</Feature>
```

Figure 1: Example of a Geonames feature's description coded in RDF

names often contain information about the type of places they are referring to, be it through the place name's surname or a keyword in the place name, sometimes that information is insufficient to distinguish the place type it corresponds to. Plus, some place types are used in various place names which have little or nothing to do with the type it refers to. Finally, people tend to make "cute" designations of places like bars or shopping malls and also use historical names for buildings that serve different purposes. All of these contribute to the mis-typing of places if based only on their names.

Spatial references or footprints may be point coordinates or bounding polygonal areas that can change place and/or coverage area over time or be too vague to locate. The quality of this representation affects its realism when referencing the place, which in turn affects the quality of the gazetteer. Therefore precision values must be established depending on the purpose of the gazetteer and also of the place the footprint represents. Dealing successfully with the issues these components bring can make the difference in the overall quality of the services a gazetteer provides.

When dealing with integrating data from multiple sources, one must consider methods to minimize time and effort in matching common aspects of their data structures. Work has been done more recently concerning the matching of feature type and export schemas of different gazetteers to both improve integration between them and add more detailed/global coverage of names [5, 6, 32]. Still, another aspect of integrating data from multiple sources concerns how to deal with multilingual data, but there aren't any guarantied methods to deal with its complexities, although an approach to tackle this would be from the perspective of the universality of spatial references and therefore trying to align the data based on comparisons between footprints.

Another aspect is the effort to try and enrich the information present in a gazetteer. A paper by Newsam and Yang describes the benefits of incorporating remote sensed imagery onto gazetteers by suggesting the use of those images as a way to improve or refine the spatial extents of the gazetteer objects [22]; another paper suggests the use of the Web to try and incorporate vernacular (common) place names and create spatial extents that fit the common perception of those places [31].

This is also referred as a problem of vague spatial extensions where the place name reflects a region that does not have a definite boundary. Instead it relies on a sense of group perception or from experience. Papers tackling this problem have proposed different types of solutions, ranging from the use of pre-defined sets of points that establish if they belong or not to some vague region [3]; use of human beings as a data source through interviews to define which known place names are considered to be in the vague region (also creates a fuzzy model by asking levels of confidence for specifying boundary locations) [21] and the use of mining techniques for information concerning known place names in the vague region from the Web [17].

4.2 Duplicate Detection

Digital gazetteers have been becoming increasingly important sources for this type of information. However, building these directories usually involve the consolidation of data from multiple data providers so as to provide the most complete information, and with it comes a important challenge: the detection and elimination or merging of exact or near duplicates for features of the same geospatial entity - Geospatial Entity Resolution.

This problem can be generalized into one of identifying database records that are syntactically different and yet describe the same physical entity and has been referred to in various ways, such as merge/purge processing [11], identity uncertainty [23], record linkage [34], among others [20, 28, 19]. The basic premise for all these methods involves computing, between pairs of records, a similarity score through the use of a similarity measure (this can be either a distance metric or a probabilistic method).

Similar records (pairs that have similarity scores higher than a given threshold) are likely to be duplicates. These can then be linked so as to form a new record, where are the pertinent information from all duplicates is merged, providing more complete and detailed version of the record. Regarding similarity measures, research has focused on the use of distance metrics computed over strings. The more commonly used metrics are:

- *Levenshtein distance* [18] (derived from the minimum number of character deletions, insertions or substitutions required to equate two strings)

- *Monge-Elkan distance* [20] (similar to the Levenshtein distance, but assigning a relatively lower cost to a sequence of insertions or deletions)
- *Jaro-Winkler metric* [34] (a fast heuristic-method specific metric for comparing proper names, which is based on the number and order of the common characters between two strings and also accounts with common prefixes)

Another form of duplicate detection involves the use of systems based on machine learning. These systems use human marked training data as a basis for learning how to classify other data. Training data can be seen as examples that illustrate relations between observed variables. In this case, the training data is in the form of record pairs that are marked as duplicates or non-duplicates by human editors. The core objective of a machine learning algorithm is to generalize from its experience, taking the training examples and extracting from them something more general, that allows it to produce useful answers in new cases.

Machine learning literature is quite extensive, so this will focus on systems that operate by training binary classifiers. Binary classification, in the machine learning domain, is the supervised learning task of classifying the members of a given set of objects (in this case, pairs of gazetteer records) into one of two groups, on the basis of whether they have some features or not. Methods proposed in the literature for learning binary classifiers include decision trees [26] and Support Vector Machines (SVMs) [16].

Decision tree classifiers learn a tree-like model in which the leaves represent classifications and branches represent the conjunctions of features that lead to those classifications. Decision tree classifiers provide high accuracy and transparency (a human can easily examine the produced rules), although they can only output binary decisions (the gazetteer record pairs would either be duplicates or non-duplicates). SVMs work by determining an hyper-plane that maximizes the total distance between itself and representative data points (i.e., the support vectors) transformed through a kernel function. SVMs can provide a measure of confidence in the result, i.e. an estimate of the probability that the assigned class is the correct one.

Regarding the context of this article, previous works have defined the problem of Geospatial Entity Resolution as the process of defining from a collection of database sources referring to locations, a single consolidated collection of true locations [29]. The problem differs from other record linkage scenarios mainly due to the presence of a continuous spatial component in geospatial data. Whereas in the case of place names, often associated with problems of ambiguity (e.g., different places may share the same name), geospatial footprints provide an unambiguous form of geo-referencing. Therefore, the problem of duplicate detection should be more simple by using geospatial data as a more precise means to join similar entities. However, in practice, spatial data is often imprecise. Different organizations often record geospatial footprints using different scales, accuracies, resolutions and structure [29]. This requires then the use of both spatial and non-spatial features, although it raises challenges due to combining semantically distinct similarity measures.

One way to combine different similarity metrics is to put a threshold on one, then using another metric as a secondary filter (i.e. helping in the rejection of similar locations according to the first metric that are not duplicates), and so on. However, the approach above does not capture the matches that are neither highly similar according to each of the individual similarity metrics. In this case, one needs a single similarity measure which combines all the individual metrics into a single score. Previous approaches have proposed to use an overall similarity between pairs of features, computed by taking a weighted average of the similarities between their individual attributes [27]. Weighted averages have the flexibility of giving some attributes more importance than others. However, tuning the individual weights can be difficult, and machine learning methods offer a more robust approach.

Another added complexity is the fact that evaluating all possible pairs of duplicate records is highly inefficient [7]. However, because most of them are clearly dissimilar non-matches, only record pairs that are loosely similar (e.g. share common tokens) can be selected as candidates for matching through the use of techniques such as blocking [11], canopy clustering [19] or filtering [35]. These three types of techniques share the fact that they explore computationally inexpensive similarity metrics in order to limit the number of comparisons that require the use of the expensive similarity metrics.

5. SYSTEM IMPLEMENTATION

Next we describe the implementation of the Gazetteer. First we present the data model, followed by a description of the system interfaces and services.

5.1 Data Representation

The data within the Gazetteer is defined using OWL for creating an internal representation of the Features. In it, a feature is a class with seven main attributes:

- ID - both an internal and source ID are stored for each feature;
- primary name associated with the feature;
- one or more alternate names, which may have a language code associated with them;
- spatial footprints or temporal coverages, depending on the feature;
- one or more feature types;
- relationships to other features (e.g. part-of, adjacency, country)
- other information (e.g. demographics, country code);

The same class is used for defining both temporal and geographic features. Time spans are associated to temporal features, spatial footprints are associated to geographic features. Geographic features must always be associated with names. However, for temporal features, the specification of time spans alone is also allowed.

In terms of the vocabulary chosen to categorize the features, an internal schema is used based on the FTT. All gazetteer features are always associated with a feature type

in this schema. In practice, it is also an OWL ontology defining classification terms and relationships among them. The footprints are defined as both GML strings representing points, bounding boxes or polygons, This representation is the one originally used in the DIGMAP project. While there have been no changes to that format, in the future, there is work planned to adjust the representation to include more attributes and possibly change some of the existing ones.

This data is stored in a relational database as records encoded in XML (each Feature class is isolated and encoded using the RDF/XML encoding schema for OWL). This enables immediate access to the complete records, eliminating the time wasted in record reconstruction. Records are also compressed prior to storage in order to optimize transfers and storage. Also, it can allow requests in different output formats, through the use of XSLT sheets for transforming the data.

An example of a feature's internal representation is shown in Figure 2.

In this figure, one can see the feature description beginning with the **Feature** node, with an attribute that provides the internal identifier for this particular feature. The primary name is provided in the **hasName** node, while any alternative names are placed inside the **hasAltName** node. In this case, there is one alternative name provided, a translation in the Occitan Language.

```
<gaz:Feature xmlns:gaz="http://www.digmap.eu/gazetteer/version1.0#"
    rdf:ID="http://sws.geonames.org/3020251/">
    <gaz:hasName>
        <gaz:Name xml:lang="">Embrun</gaz:Name>
    </gaz:hasName>
    <gaz:hasAltName>
        <gaz:Name xml:lang="oc">Ambrun</gaz:Name>
    </gaz:hasAltName>
    <rdfs:type xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
            rdf:resource="http://www.esri.com/metadata/catalog/adl/#populated
    <gaz:altType rdf:resource="http://www.geonames.org/ontology#P"/>
    <gaz:hasCode rdf:resource="http://www.geonames.org/ontology#P.PPL"/>
    <gaz:countryCode>FR</gaz:countryCode>
    <gaz:population>7069</gaz:population>
    <gaz:postalCode>05200</gaz:postalCode>
    <gaz:postalCode>05209</gaz:postalCode>
    <gaz:postalCode>05201</gaz:postalCode>
    <gaz:postalCode>05208</gaz:postalCode>
    <gaz:postalCode>05202</gaz:postalCode>
    <gml:centerOf xmlns:gml="http://www.opengis.net/gml">
        <gml:Point>
            <gml:coord><gml:X>44.56387</gml:X><gml:Y>6.49526</gml:Y></gml:coo
        </gml:Point>
    </gml:centerOf><boundedBy xmlns="http://www.opengis.net/gml">
    <Envelope>
        <lowerCorner>
            <coord><X>44.5638722847772</X><Y>6.49525880813599</Y></coord>
        </lowerCorner>
        <upperCorner>
            <coord><X>44.5638722847772</X><Y>6.49525880813599</Y></coord>
        </upperCorner>
    </Envelope></boundedBy>
    <gaz:hasFootprint>
        <gaz:Footprint gaz:primary="true">
            <Point xmlns="http://www.opengis.net/gml">
                <coord><X>44.5638722847772</X><Y>6.49525880813599</Y></coord>
            </Point>
            <gaz:CSquares>1004:364:245</gaz:CSquares>
            <gaz:GeoHash>sputeb9eqebh</gaz:GeoHash>
            <gaz:WKT>POINT (44.5638722847772 6.49525880813599)</gaz:WKT>
        </gaz:Footprint>
    </gaz:hasFootprint>
    <gaz:partOf rdf:resource="http://sws.geonames.org/6446638/"/>
    <gaz:inCountry rdf:resource="http://sws.geonames.org/3017382/"/>
```

Figure 2: Excerpt of internal Gazetteer Feature of Embrun

Following is some descriptive information about the feature, namely its country code, population and postal codes. Then, there are some geographic descriptions of the feature: point coordinates, bounding box representation, all according to the GML vocabulary. Finally, relationships this feature shares with others are described.

5.2 Human Interface

The Gazetteer Public User interface serves as the portal for human use of the gazetteer. In it, a user can search for features and, through the results, browse to other features, through their classes and relationships.

From its main page, a user can perform a search for a feature name, for which any matching features are display in a result list. In the result list, one can already visualize, for each search match, their:

- Feature name;
- Feature type;
- Graphic approx. location;
- Provenience;

From this point, a user can either choose one of the results to browse within or perform another search. When choosing one of the features, the user is presented with a description page. There is a descriptive information portion of the page where information regarding relationships can be accessed, so they can be used as a means to browse to other features and, if existing, information articles can be accessed to offer other data not present in the gazetteer or to offer descriptions in other languages. Another portion shows the geographic information, where one can see both a visual representation of the feature and coordinate representations;

Below all this information is a board where the feature has been defined in several data types, such as the ADL Content Standard, Geonames Ontology or KML.

5.3 ADL-GP interface

The gazetteer can be accessed via an XML over HTTP request interface and the request/response protocol is the ADL-GP protocol. Requests may be invoked using HTTP POST requests to the gazetteer URL specified below with the required parameters, and all text encoded in UTF-8. The current deployment of the Gazetteer, for demonstration purposes, has the base URL: `http://europeana-geo.isti.cnr.it/gazetteer/services/gp` - this URL serves as the entry point for incoming XML requests. The base format for every request is shown in FIgure 3. The query type(s) can then be specified inside the **query-request** node. Following are some types of queries the system supports and examples.

```
<?xml version="1.0" encoding="UTF-8"?>
<gazetteer-service
    xmlns="http://www.alexandria.ucsb.edu/gazetteer"
    version="1.2">
  <query-request>
    <gazetteer-query>
      ...
    </gazetteer-query>
    <report-format>standard</report-format>
  </query-request>
</gazetteer-service>
```

Figure 3: Base format for a gazetteer feature query

A: *Identifier Query*

Description: This simple query type fetches the feature, if existing, which matches with the provided internal identifier as argument.

Example:
```
<identifier-query
identifier="http://sws.geonames.org/299042/" />
```

B: *Name Query*

Description: This query type fetches the features whose name, being it primary or alternative, matches with the provided text as argument, depending on the matching operator provided as argument. Those different types of operators may be:

- **"*equals*":** A name, in its entirety, matches the exact text;
- **"*contains-all-words*":** A name must contain all words specified in the text, in no particular order;
- **"*contains-any-words*":** A name must contain at least one of the words specified in the text;
- **"*contains-phrase*":** A name must contain the exact sequence of words specified in the text;

Example:
```
<name-query
operator="equals" text="lisboa" />
```

C: *Class Query*

Description: This query type fetches the features whose classification matches the term provided. The thesaurus used for search purposes is the ADL Feature Type Thesaurus.

Example:
```
<class-query
thesaurus="ADL Feature Type Thesaurus"
term="streams" />
```

E: *Combine Queries*

Description: Besides using each of the previous query types by itself, it is also possible to combine different query types with one another in order to form more complex queries, though the use of an ***and*** clause.

Example:
```
<and>
  <class-query
   thesaurus="ADL Feature Type Thesaurus"
   term="streams" />
  <name-query
   operator="equals" text="Danube" />
</and>
```

5.3.1 Query Response

The XML response for any query request type consists of a list of all search matches to the query parameters. Figure 3 depicts the base format of a query response:

For each standard report entry, there are these main attributes described:

- The feature's internal identifier (*adlgp:identifier*), which corresponds to the source providerÕs identifier;
- The display name (*adlgp:display-name*), which is the primary name associated with the feature;
- A list of alternative names (*adlgp:names*), where each one may have a language code, corresponding to the source language for that name;
- A country code (*gaz:countryCode*), as defined by the ISO 3166[7], that identifies the pertaining country of that feature;
- Geographic representations of the feature (Point, Bounding Box, ..);
- The feature types that classify the feature (*adlgp:classes*):
  - For each one, there is an attribute that determines if that class is the primary type and another attribute that identifies the thesaurus of origin for that class;
- A list of relationships with other features (*adlgp:relationships*):
  - For each one, there is an attribute that identifies the relationship type and another that identifies the target feature identifier;
- Some other elements may be present, depending on the feature's provenience, such as Postal Codes, Population, external information articles (Wikipedia , DBpedia), etc.

An example is shown in Figure **??**.

5.4 SRU Interface

The SRU interface serves as another form for querying content in the gazetteer, as in the case of the ADL-GP interface. The protocol to which it complies - Search/Retrieval via URL (SRU)[8] - is a standard search protocol for queries on the Internet, utilizing Contextual Query Language (CQL), a standard syntax for representing queries. It's relative ease of implementation and use has made it popular in new applications for search purposes.

The current deployment of this interface, for demonstration purposes, has the base URL: `http://digmap3.ist.utl.pt:8080/gazetteer-webapp/services/sru` - this URL serves as the entry point for the SRU protocol and presents the explain record for the deployed version of the protocol. From there, one constructs requests by editing the URL as established by the CQL syntax.

5.5 Ingestion Service

The Ingestion Service is an external component designed exclusively for importing and updating content from data providers. The main goal is to provide a component that can take a data provider's content and process it in order to convert the content to the gazetteer's format. Because different data providers use different forms of storing their content and also different data formats, there is no standard and unique way to obtain content. Therefore, the service provides different processing methods depending on the chosen data provider.

[7] `http://www.iso.org/iso/country_codes.htm`

[8] A detailed explanation of the service is provided in chapter `http://www.loc.gov/standards/sru/`

```
<?xml version="1.0" encoding="UTF-8"?>
<gazetteer-service xmlns="http://www.alexandria.ucsb.edu/gazetteer">
    <query-response>
        <standard-reports>
            <adlgp:gazetteer-standard-report xmlns:adlgp="http://www.alexandria.ucsb.edu/gazetteer"
                                    xmlns:fn="http://www.w3.org/2005/02/xpath-functions"
                                    xmlns:gaz="http://www.digmap.eu/gazetteer/version1.0#"
                                    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
                ....
            </adlgp:gazetteer-standard-report>
            ....
        </standard-reports>
    </query-response>
</gazetteer-service>
```

Figure 4: Base format for a gazetteer feature query response

At the moment, the service supports importing content from Geonames (by importing a RDF data dump) and from Geo-Net-PT 02[9] (also by importing a RDF data dump). For both sources, the required data is provided as RDF files and is processed in the following manner:

1. The content is read from the data file and parsed as multiple features, each one added as a whole to a new database table representing the data source.
    - Each record is stored in a row with two fields: an "ID" field which stores the data providers' ID for the particular feature and a second field "Content" which stores the entire structured RDF text of the Record.
2. Then, content is transformed into the gazetteer's own internal format, through SQL operations and XSLT stylesheets, which discards unnecessary fields and merges others
3. At this point, the duplicate detection service is invoked in order to allow filtering the content from the data source for duplicate records

The result is a new table with the data source content transformed into the gazetteer internal format ready for querying. A considered way to expand the service is to couple it with another external service - REPOX[10] - a data aggregator that exposes its stored data via Open Archives Initiative - Protocol for Metadata Harvesting (OAI-PMH)[11], so as to delegate responsibility to REPOX for checking sources for updated data.

5.6 Duplicate Detection Service

The Duplicate Detection Service is an external component designed exclusively for generating possible pairs of duplicates in a feature collection. The goal is to take a collection of geographic features and generate a list of possible duplicates. This list can then be used to filter the collection from unnecessary features. The service should provide more than one technique for duplicate detection, to allow choosing the appropriate method depending on the precision required. At the moment, however, only one method is available for processing feature collections for duplicates.

The available method combines two similarity metrics. It compares primarily by feature name, using a Jaro-Winkler distance metric. This metric seemed more appropriate since it is specific for comparing proper names. A minimum threshold score is used to get a collection of possible duplicate candidates. These pairs are then compared by geographic distance between features, that is, the distance between centroids of the two features is measured and, if below a certain threshold, are considered valid candidate pairs. These threshold scores were obtained through tests with the feature collections that were imported and are described in the following chapter.

This comprises the comparison method. However, as it was explained in the Main Concerns Section, testing all possible combinations of duplicate pairs is highly inefficient. Also, because the service can deal with large volume collections of features (just the Geonames dataset comprises over 7 million features, with a size of over 10 Gigabytes), it is practical, if not necessary to reduce the amount of feature comparisons to actually perform. In that case, before performing any sort of comparison between features, the feature collection is first clustered into batches which are organized by two criteria: first by country of origin (similar features are gathered in a same geographic region), then by the type of feature it consists of. In this way, the amount of feature comparisons is clearly reduced, by taking only record pairs with similar attributes into account.

With the detection process explained, we can now describe the execution flow for the duplicate detection service, when processing a collection. To note, the input of the system is a database table of gazetteer features, which are expected to be converted to the internal format:

1. From an initial database table, another table is created containing the feature identifiers, name, country and feature type.
2. Using this newly created table, batches of features are partitioned by those from the same country and subsequently, similar type.
3. Then, for each batch, comparisons are made between each possible pair of features for string similarity regarding the primary feature names and, then, by geographic distance.

[9] http://xldb.fc.ul.pt/wiki/Geo-Net-PT_02_in_English
[10] http://repox.ist.utl.pt/
[11] http://www.openarchives.org/pmh/

- The name comparisons are made through use of an implementation of the Jaro-Winkler similarity metric, from SecondString[12], an open-source Java based collection of approximate string matching techniques.
- The distance measure is calculated though the use of the spherical law of cosines, which is a theorem relating the sides and angles of spherical triangles, analogous to the ordinary law of cosines from plane trigonometry, where R is the Earth's Radius with the value of 6,371 km.:

$$d = \arccos(\sin(lat1) * \sin(lat2) + \cos(lat1) * \cos(lat2) * \cos(long2 - long1)) * R$$

4. The result are lists for each country containing candidate duplicates. These can then be either used to merge contained pairs or browse through them for confirmation.

5.7 Data Statistics and Classifying

In order to test ingestion mechanism, a new database instance for the gazetteer was created by importing two data sources: Geonames and Geo-Net-PT 02[13]. This section the statistics regarding the two data-sources imported and statistics for the end result internal database.

Table 1 shows the statistics regarding the content from the Geonames RDF dump. Table 2 shows statistics regarding the content from GeoNet-PT 02 RDF dump.

Names - Total	10217972
Names - Average p/ Feature	1.366
Names - Unique	6833275
Countries - Total	251
Features - Average p/Country	29779
Features - No Country assoc.	5391
Countries - Least Features	NF (5)
Countries - Most Features	US (2059852)
Feature Types - Total	618
Features - No Type assoc.	5502
Features - Average p/Type	12103
Features - Total	7479708
Features - Dataset size	10140 MB

Table 1: Geonames data statistics

Features - Place Names	270816
Features - Feature Types	89
Features - Footprints	7804
Features - Geographic Features	204729
Features - Total	724565
Features - Dataset size	580 MB

Table 2: Geo-Net-PT 02 data statistics

[12]http://secondstring.sourceforge.net/

[13]These collections were obtained, respectively, via the URLs `http://download.geonames.org/all-geonames-rdf.zip` and `http://www.linguateca.pt/geonetpt/geonetpt02/`

After the ingestion process, each feature is assigned a primary feature type belonging to the ADL Feature Type Thesaurus and maintains his original feature type as a secondary type. For the purposes of these statistics, the feature types considered will be the ADL ones. Table 4 shows the resulting database statistics.

Features - Place Names	7026118
Features - Feature Types	618
Features - Footprints	7487512
Features - Relationships	15398856
Features - Relationship Types	4
Features - Total	7684437
Nij DataSets	2
Features - Size	12324 MB

Table 3: Gazetteer data statistics

The last line regarding the database size refers to just the converted features, however the database still keeps a table for each imported dataset in its original form, so as to facilitate checking for differences when updating a particular dataset.

Next, we provide some statistics so as to demonstrate the testing with the detection process in order to achieve a lower margin of "false positives" (excess of duplicate candidates). The main goal in developing the technique for the system was to reduce the number of comparisons between features. This was both due to the amount of data the service would have to process and the fact that testing all possible combinations of duplicate pairs is highly inefficient, as explained earlier.

To accomplish this, the first phase in the duplicate detection system is clustering the features into batches which are organized by two criteria: primarily by country of origin and then by feature type.

Then, for each batch, comparisons are made between all possible pairs of features' primary names. Finally, the remaining candidates are compared by distance between their coordinates, being selected if their distance is below a certain threshold, provided they exist for each candidate pair. For this test, the considered threshold was of 15Km.

The testing was made using a selection of features from the Gazetteer database, grouped by pertaining country. Then, the selection was evaluated by the Duplicate Detection system, producing statistics regarding the number of possible duplicate pair candidates found, their distance threshold in Km. The following table shows a summary of the results from that test.

Batch Total Features	553552
Found Duplicate Candidates	1164
Duplicates with Distance Margin < 1Km	40

Table 4: Duplicate Detection data statistics

After testing a batch of 53552 features, pertaining to 20 different countries, a total of 1164 duplicate candidates were proposed (close to 2,2% of the batch size). Of those, 40 (3%

of the duplicates) were determined to be close to each other below a threshold of 1Km. This, in the context of comparing two populated areas is cause for seriously considering they are referring to the same location.

5.8 Results and Conclusions

For the purposes of this report, this work's goal was to analyze a service (Gazetteer) made for a previous project (DIGMAP) and adapt that system in order to adjust to the needs of another project (EuropeanaConnect).

The main requirement for the service was to serve as a knowledge database for another service developed for the project - the Geoparser. Still, during the course of the work for the project, other requirements and goals were established, so as to establish an infrastructure of components and external services that interacted with the Gazetteer service, enhancing its functionality and performance.

In concrete terms, we have implemented the main gazetteer service providing three forms of access - Human, ADL-GP and SRU interfaces and around it, other services that were designed to enhance its performance - a content Ingester and a duplicate detector. While the end result currently meets the intended requirements for the target project, there is still a difference in terms of proposed requirements/accomplished requirements:

- While the planned solution detailed a collection of three external services that served to expand functionality in the gazetteer, only two were completed at this point - the Ingester and the Duplicate Detector;
- Both the Ingester and Duplicate Detector are in a basic state of operation, importing only a restricted collection of data sources and only providing one basic technique for duplicate detection;
- An administrative interface for managing the gazetteer content and use the external services is still not operational.
- An addition to the end result that was not planned was the inclusion of another querying interface - the SRU protocol. This was added at a final stage by means of a suggestion made through reviews of the Gazetteer Service for the Europeana Connect project.

I believe the main difficulties in working in this project arose with the amount of data to work with (databases and text files in order of Gigabytes), which would have to be processed with in terms of data quality. Overall, despite the differences, the main goal was set, as the service was accepted and positively reviewed.

In closing, after researching about gazetteers, one can have a better understanding about their concept, role and representative projects concerning them. Also, it allows one to have a better understanding of the problems that they have, their causes and attempts at solutions. Mainly, this report serves as a starting point for developing new solutions and expanding current ones for gazetteer use, by providing a better insight of the problem at hand, its complexities and potential approaches.

6. REFERENCES

[1] ISO/IEC IS 19101:2002: Geographic Information - Reference Model. International Organization for Standardization, Geneva, Switzerland.

[2] ISO/IEC IS 19112:2003: Geographic Information - Spatial Referencing by Geographic Identifiers. International Organization for Standardization, Geneva, Switzerland.

[3] H. Alani, C. B. Jones, and D. Tudhope. Voronoi-based region approximation for geographical information retrieval with gazetteers. *International Journal of Geographical Information Science*, 15(4):287–306, 2001.

[4] J. Borbinha, G. Pedrosa, J. Luzio, H. Manguinhas, and B. Martins. The DIGMAP virtual digital library. *e-Perimetron*, 4(1):1–8, 2009.

[5] D. F. Brauner, M. A. Casanova, and R. L. Milidiú. Towards gazetteer integration through an instance-based thesauri mapping approach. In *Proceedings of the 8th Brazilian Symposium on GeoInformatics*, 2006.

[6] D. F. Brauner, C. Intrator, J. C. Freitas, and M. A. Casanova. An instance-based approach for matching export schemas of geographical database web services. In *IX Brazilian Symposium on GeoInformatics*, 2007.

[7] A. Elmagarmid, P. Ipeirotis, and V. Verykios. Duplicate record detection: A survey. *IEEE Transactions on knowledge and data engineering*, pages 1–16, 2007.

[8] M. F. Goodchild and L. L. Hill. Introduction to digital gazetteer research. *International Journal of Geographical Information Science*, 22(10):1039–1044, 2008.

[9] J. Hastings and L. Hill. Treatment of duplicates in the alexandria digital library gazetteer. In *GIScience 2002*, 2002.

[10] J. T. Hastings. Automated conflation of digital gazetteer data. *International Journal of Geographical Information Science*, 22(10):1109–1127, 2008.

[11] M. A. Hernandez and S. J. Stolfo. The merge/purge problem for large databases. In *Proceedings of the 1995 ACM Conference on Management of Data.*, 1995.

[12] L. Hill, J. Frew, and Q. Zheng. Geographic names: The implementation of a gazetteer in a georeferenced digital library. *D-Lib Magazine*, 1999.

[13] L. L. Hill. Core elements of digital gazetters: Placenames, categories, and footprints. *Lecture Notes in Computer Science*, pages 280–290, 2000.

[14] L. L. Hill and Q. Zheng. 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*, volume 36, pages 57–69, 1999.

[15] K. Janowicz and C. Keßler. The role of ontology in improving gazetteer interaction. *International Journal of Geographical Information Science*, 22(10):1129, 2008.

[16] T. Joachims. Making large-scale svm learning practical. *Advances in Kernel Methods Support Vector*

Learning, pages 169–184, 1999.

[17] C. B. Jones, R. S. Purves, P. D. Clough, and H. Joho. Modelling vague places with knowledge from the web. *International Journal of Geographical Information Science*, 22(10):1045–1066, 2008.

[18] V. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet Physics Doklady*, volume 10, pages 707–710, 1966.

[19] A. McCallum, K. Nigam, and L. Ungar. Efficient clustering of high-dimensional data sets with application to reference matching. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 169–178. ACM, 2000.

[20] A. Monge and C. Elkan. The field matching problem: Algorithms and applications. In *Proceedings of the second international Conference on Knowledge Discovery and Data Mining*, pages 267–270, 1996.

[21] D. R. Montello, M. F. Goodchild, J. Gottsegen, and P. Fohl. Where's downtown?: Behavioral methods for determining referents of vague spatial queries. *Spatial Cognition and Computation*, 3(2&3):185–204, 2003.

[22] S. Newsam and Y. Yang. Integrating gazetteers and remote sensed imagery. In *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems*, page 26, 2008.

[23] H. Pasula, B. Marthi, B. Milch, S. Russell, and I. Shpitser. Identity uncertainty and citation matching. *Advances in Neural Information Processing Systems*, (15):1425–1432, 2003.

[24] E. Rauch, M. Bukatin, and K. Baker. A confidence-based framework for disambiguating geographic terms. In *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references-Volume 1*, page 54, 2003.

[25] J. Ressler, E. Freese, and V. Boaten. Semantic method of conflation. *Proceedings of the Terra Cognita Workshop, collocated with the 8th International Semantic Web Conference (ISWC-2009), Washington DC, USA*, 518, Oct. 2009.

[26] S. Safavian and D. Landgrebe. A survey of decision tree classifier methodology. *Systems, Man and Cybernetics, IEEE Transactions on*, 21(3):660–674, 1991.

[27] A. Samal, S. Seth, and K. Cueto. A feature-based approach to conflation of geospatial sources. *International Journal of Geographical Information Science*, 18(5):459–489, 2004.

[28] S. Sarawagi and A. Bhamidipaty. Interactive deduplication using active learning. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, page 278. ACM, 2002.

[29] V. Sehgal, L. Getoor, and P. Viechnicki. Entity resolution in geospatial data integration. In *Proceedings of the 14th annual ACM international symposium on Advances in geographic information systems*, pages 83–90. ACM, 2006.

[30] D. A. Smith and G. Crane. Disambiguating geographic names in a historical digital library. *Lecture Notes in Computer Science*, pages 127–136, 2001.

[31] F. A. Twaroch, C. B. Jones, and A. I. Abdelmoty. Acquisition of a vernacular gazetteer from web sources. In *Proceedings of the first international workshop on Location and the web*, pages 61–64, 2008.

[32] O. Vestavik and I. T. Solvberg. Merging local and global gazetteers. *Lecture Notes in Computer Science*, 4822:495, 2007.

[33] R. Volz, J. Kleb, and W. Mueller. Towards ontology-based disambiguation of geographical identifiers. In *16th International World Wide Web Conference (WWW2007), Banff, Alberta, Canada*, 2007.

[34] W. Winkler, W. Winkler, et al. Overview of record linkage and current research directions. In *Bureau of the Census*. Citeseer, 2006.

[35] C. Xiao, W. Wang, X. Lin, and J. Yu. Efficient similarity joins for near duplicate detection. In *Proceeding of the 17th international conference on World Wide Web*, pages 131–140. ACM, 2008.

7. APPENDIX - GLOSSARY

ADL Alexandria Digital Library

DIGMAP DIscovering Our Past World with Digitised Maps

ECAI Electronic Cultural Atlas Initiative

ESRI Environmental Systems Research Institute

FTT Feature Type Thesaurus

GCS Gazetteer Content Standard

GIS Geographic Information Systems

GML Geography Markup Language

GNIS Geographic Names Information System

ISO/TC 211 International Organization for Standardization Technical Committee 211

NIMA National Imagery and Mapping AgencyÇ

OGC Open Geospatial Consortium

OWL OWL Web Ontology Language

RDF Resource Description Framework

TGN Getty Thesaurus of Geographic Names

URL Uniform Resource Locator

VGI Volunteered Geographic Information