INSTITUTO SUPERIOR TÉCNICO
Universidade Técnica de Lisboa

Design and implementation of a Gazetteer

André Soares

Dissertação para obtenção do Grau de Mestre em
Engenharia Informática e de Computadores

Júri

Presidente:	Mário Rui Gomes
Orientador:	José Borbinha
Co-Orientador:	Bruno Martins
Arguente:	Gabriel Pestana

Julho 2011

Agradecimentos

Agradeço aos meus orientadores pela paciência e apoio e à minha família pelos trabalhos causados.

Lisboa, October 12, 2011
André Soares

Dedico isto aos meus avós, já falecidos, pelo
papel que tiveram na minha vida.

Resumo

O objectivo deste trabalho consiste na análise de um servico Gazetteer obtido de um projecto anterior (DIGMAP) para depois adaptar com base nos requisitos de um outro projecto Europeu (Europeana Connect). O produto final é um servico gestor de conteúdo geográfico que interage com outra ferramenta - um Geoparser. O resultado foi um servico básico que, embora cumpre os requisitos estabelecidos pela Europeana Connect, não dispõe ainda de muita funcionalidade acrescida, sendo apenas capaz de fazer ingestão de conteúdo de um número limitado de "data providers" e dispõe de técnicas limitadas para a detecção de duplicados. Durante o curso do relatório, dá-se a descrição do conceito de um Gazetteer e a sua evolução, juntamente com outros conceitos relevantes. Depois segue-se um estudo dos projectos mais relevantes que contribuíram para o desenvolvimento dos Gazetteers, seguido de uma descrição das áreas mais relevantes de estudo actualmente. Depois, apresenta-se o problema do trabalho em si, juntamente com os seus requisito. Segue-se a proposta da arquitectura do serviço e a explicação do que se encontra actualmente implementado. Finalmente, apresenta-se algumas estatísticas relativamente aos dados importados e uma medição da performance do serviço para detecção de duplicados, terminando com um sumário dos objectivos concretizados, uma explicação dos que faltaram, com as razões possíveis para as diferenças e um plano de trabalhos futuro.

Palavras-chave: Gazetteer, GIS, Features, Europeana

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, personal conclusions and draw plans for future work.

Keywords: Gazetteer, Geographic Features, Data Quality, Europeana Connect

Contents

List of Figures

List of Tables

Acronyms

ADL	Alexandria Digital Library
FTT	ADL Feature Type Thesaurus
GCS	ADL Gazetteer Content Standard
GIS	Geographic Information Systems
GML	Geography Markup Language
HTML	Hypertext Markup Language
ISO	International Organization for Standardization
LOD	Linking Open Data
OGC	Open Geospatial Consortium
OWL	Web Ontology Language
RDF	Resource Description Framework
TGN	Getty Thesaurus of Geographic Names
URI	Uniform Resource Identifier
VGI	Volunteered Geographic Information
XML	Extensible Markup Language
XSLT	Extensible Stylesheet Language Transformations

Chapter 1

Introduction

1.1 Context

Computers have opened a vast new potential in the way we communicate, analyze our surroundings, and make decisions. Data representing the geographical aspects of our world can nowadays be stored and processed efficiently, and many of the decisions people make today depend on the details of our immediate surroundings, requiring information about specific places on the Earth's surface. This information, labeled *geographical information* helps distinguish places and helps to make decisions pertaining to specific locations[10]. As such, specialized computer systems have been developed for the acquisition and manipulation of geographical information in various ways. These systems are known as Geographic Information Systems (GIS). The kind of GIS that's the target of this dissertation is that of a gazetteer.

A gazetteer consists of a list of geographic names, together with their geographic locations and other descriptive information [32]. 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. In the effort to enhance that role, organizations such as ISO and OGC are vital elements because they attempt to establish standardization in the digital aspect of geography (geographic information systems, data representations, etc.).

1.2 Motivation

The motivation for this work was to construct a new design, implementation and validation of a gazetteer initially developed for the DIGMAP project [11], 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].

The objective was the construction of a new improved version of a Gazetteer service and at the same time improve upon the current standard for a Gazetteer service.

[1] http://www.flickr.com

[2] http://wikimapia.org

[3] http://www.europeanaconnect.eu

[4] http://www.europeana.eu

1.3 Problem

The Europeana service has the goal to enable people to explore the digital resources of Europe's museums, libraries, archives and audio-visual collections. Given the extent and diversity of the resources contained, there should be several ways to relate content between them and allow searches for it.

In that sense, one of the goals of the Europeana Connect project was the integration of value-adding services that would further enhance Europeana's functionality. One of these services is a GIS suite designed to allow users to query and display content based on a spatial dimension, or to discover new relationships between content based on location.

In that context, the Gazetteer is a required component for the GIS suite to be delivered.

The result of this work is a new version of both the technological solution of the former Gazetteer and its associated service. It had to comply with the requirements set by Europeana Connect, while managing to deliver both according to the project timeframe and Europeana's release schedule. Also, the results are production-ready components.

1.4 Results and Contributions

The results gathered from this work show that while the entire planned system is still incomplete, it 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 resulting Gazetteer System is comprised by two parts: first there's the main International Organization for Standardization (ISO) and Open Geospatial Consortium (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 added services include 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 service ecology, which serves as a sort of "back-office" for the standard Gazetteer service, in order to improve its performance and reliability.

Further details and descriptions can be seen in this page `http://gaz.ist.utl.pt`, dedicated to the project, describing its motivation, goals, features and with documentation and links to testing interfaces.

1.5 Structure of the document

The remainder of this document is structured as follows. First, a look at the current state of the art in gazetteers is presented. This begins with a more detailed description of a Gazetteer, according to current best practices, along with an account of the ongoing efforts for standardization of a Gazetteer Service.

Then ensuing an account of some projects/initiatives that contributed for its development, while at the same time giving some examples of gazetteer applications. This is followed by a look into the active areas of research and development, with discussion of their main concerns, challenges and goals in order to improve gazetteers. Also, a look into the state of the art in duplicate detection, specifically more targeted for the problem of gazetteer records and also the Linking Open Data (LOD) initiative.

Next, a definition of the problem: redesign and re-implementation of a pre-existing gazetteer, adapting the solution to respond towards the necessities of the Europeana project. A detailed account of the work process follows, along with a description of the innovative aspects incorporated into the new gazetteer (duplicate detection for better data quality and publication of the data through the LOD[5] principles).

Finally, a description of the validations for the developed solution along with the obtained results, the conclusions based upon those results and future work planning.

[5] `http://linkeddata.org`

Chapter 2

State of the art

This chapter provides the technical context for this dissertation. In it, the concept of a gazetteer is presented in detail, along with any relevant terminology for the purposes of understanding the work. Then, a summary of the more relevant projects and initiatives made in terms of gazetteers and for the purposes of this dissertation is provided.

2.1 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 and the OGC, have led us to a point where we have a general agreement of their basic structure, attributes and basic protocols.

The concept presented in this dissertation combines ISO standards with OGC specifications and discussion documents. This definition includes the attributes, properties and minimum requirements that have been established so far for a gazetteer system. An important aspect of gazetteers are the entities they store, named features [1] :

Definition 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, features are distinct physical elements or objects (buildings, rivers, mountains) for which geographic positions and boundaries can be established [23]. 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 [4]. In this context:

Definition 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).*

A 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 [4] can now be introduced :

Definition 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. Figure 2.1 shows an example of a class diagram of a gazetteer model according to the previous definitions established [19].

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* [34]. 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 [58, 61, 51].

2.2 Standardization Efforts

ISO/TC 211 and the OGC emerged when there was need for standardization efforts in the areas of geography and cartography, but the initial ones were slow and difficult. National and international organizations were busy developing standards for the transfer or exchange of geographic data between computer systems.

However, the technical development of such standards was limited to few national and regional user communities and there were no standards that had broad international support. With the emerging of these international standardization efforts, they attempted to gain the international recognition and

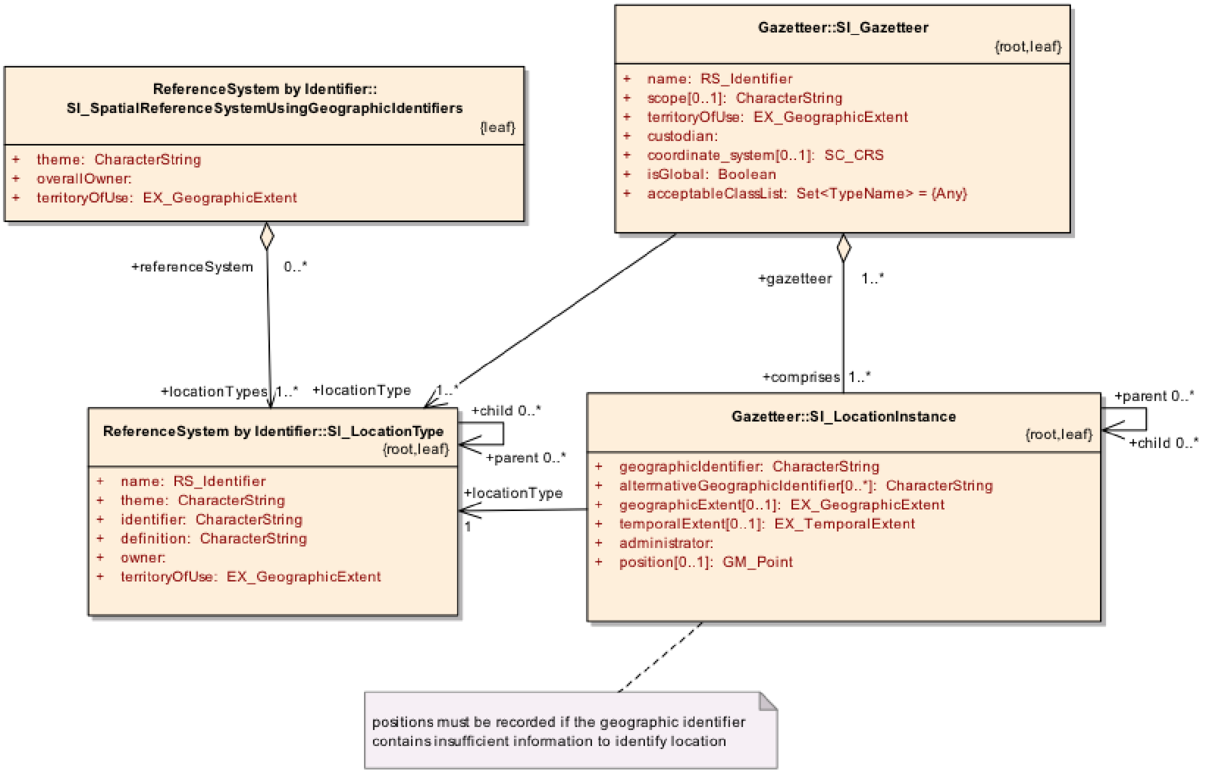

Gazetteer::SI_Gazetteer
{root,leaf}
+ name: RS_Identifier
+ scope[0..1]: CharacterString
+ territoryOfUse: EX_GeographicExtent
+ custodian:
+ coordinate_system[0..1]: SC_CRS
+ isGlobal: Boolean
+ acceptableClassList: Set<TypeName> = {Any}
ReferenceSystem by Identifier::
SI_SpatialReferenceSystemUsingGeographicIdentifiers
{leaf}
+ theme: CharacterString
+ overallOwner:
+ territoryOfUse: EX_GeographicExtent
+referenceSystem
0..*
+gazetteer
1..*
+comprises
1..*
+locationTypes
1..*
+locationType
1..*
+child 0..*
+parent 0..*
ReferenceSystem by Identifier::SI_LocationType
{root,leaf}
+ name: RS_Identifier
+ theme: CharacterString
+ identifier: CharacterString
+ definition: CharacterString
+ owner:
+ territoryOfUse: EX_GeographicExtent
+locationType
1
Gazetteer::SI_LocationInstance
{root,leaf}
+ geographicIdentifier: CharacterString
+ alternativeGeographicIdentifier[0..*]: CharacterString
+ geographicExtent[0..1]: EX_GeographicExtent
+ temporalExtent[0..1]: EX_TemporalExtent
+ administrator:
+ position[0..1]: GM_Point
+parent 0..*
+child 0..*
positions must be recorded if the geographic identifier contains insufficient information to identify location

Figure 2.1: Gazetteer class diagram (adapted from ISO 19112)

acceptance by the cartographic and geographic communities of the need and value of geographic standardization. In order to minimize technical overlap and manage mutual development, the ISO TC 211 and OGC also share a cooperative agreement. Under that agreement, there is an ISO/OGC Joint Advisory Group to discuss the relationship and make recommendations regarding potential new work items[1].

In general, OGC develops software interface specifications, while ISO/TC 211 develops geographic data standards. The technical committee's objectives include increasing the understanding and usage of geographic information and also increase its availability, access, integration, and sharing. For the most part, the ISO 19100 series of standards is about modeling the various aspects of the geospatial information domain, such as an abstract model for geometry, for features, for navigation, and so forth. ISO 19107 and 19108 relate to the definition of spatial and temporal schemas, respectively. ISO 19111 and 19112 concern spatial referencing systems by use of coordinates and geographic identifiers (ISO 19112 also establishes the concept of a gazetteer) and 19115 refers to metadata [4, 3, 2]. These are all examples of ISO published standards.

The OGC is an international voluntary consensus organization comprised of members from government, the commercial sector, non-governmental organizations and academia. The purpose of the OGC is to serve as a global forum for the collaboration of developers and users of spatial data products and services, and to advance the development of international interface and encoding standards that can be implemented for geospatial software interoperability. They call their produced documents Implementation Specifications. To accomplish this, they set themselves to provide free, openly available specifications, lead in their creation and establishment and facilitate their adoption worldwide. Because of this, these specifications, which collectively form a reference architecture for interoperability, have been implemented in hundreds of commercial and open source geoprocessing products and are being implemented in communities and organizations around the world.

[1] Taken from `http://www.isotc211.org/Agreements/Agreement_OGC.pdf`

Some of the more relevant specifications include the Geography Markup Language (GML)[2] which is their designed grammar for expressing geographical features. As such, it is the utilized modeling language for their interface specifications like the Web Mapping Service (WMS) or the Web Feature Service (WFS) which provide an interface allowing requests for map images/geographical features across the web using platform-independent calls [15, 62].

Also there is the Simple Features Access Specification, which is designed to both increase interoperability between WFS servers and to improve the ease of implementation of the WFS specification [31]. WFS is also the service from which the OGC is defining their Gazetteer Service, as a specialized profile of the WFS service, using a re-factored content model based on the ISO 19112 one. However, this is still in the form of a discussion paper [19]. Much of the content of their specifications (in terms of the underlying models) are based on the ISO documents.

2.3 Related Initiatives

In this section we will describe 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.

2.3.1 Getty Thesaurus of Geographic Names

The Getty Thesaurus of Geographic Names (TGN) [25] is one of the vocabularies produced by the Getty Vocabulary Program of the J. Paul Getty Trust[3] along with the Art and Architecture Thesaurus (AAT) and the Union List of Artist Names (ULAN). These are structured vocabularies designed to provide terminology and additional information about art, architecture and material culture; in the case of the TGN, it provides with names and associated information about places important to various disciplines that specialize in the subjects aforementioned.

Some of their applications include the use as data standards for the purpose of documentation cataloguing; serving as a controlled vocabulary; use as search assistants for retrieval systems and as research tools, because of their nature as a rich information source. Their main audience consists of people and institutions related to art, although a number of users of the vocabularies are students and members of the general public.

Work on the TGN began in 1987, when the trust created a department devoted to compiling and subsequent releasing of terminology (the current Vocabulary Program). Work was already ongoing in the development of the AAT and they attempted to respond to ongoing requests from creators of art information for more controlled vocabularies regarding geographic names and artist names (ULAN). The need for the vocabulary was corroborated by an international study completed by a working group of the Comité International d'Histoire de l'Art (CIHA), and by the consensus reached at a colloquium attended by the spectrum of potential users of geographic vocabulary in processes of art, architecture and archaeology.

Since then, TGN has been constructed over the years by numerous members of the user community and an army of dedicated editors, under the supervision of several managers. It was first published in 1997 but given its growing size and frequent changes, it is currently only published via Web search interfaces and data files (via licensing). The principles applied for constructing and maintaining the TGN are the same for the other vocabularies as well:

[2] http://www.opengeospatial.org/standards/gml

[3] http://www.getty.edu

- scope includes terminology needed to catalog and retrieve information about the visual arts and architecture;

- uses national [7] and international [35] standards for thesaurus construction;

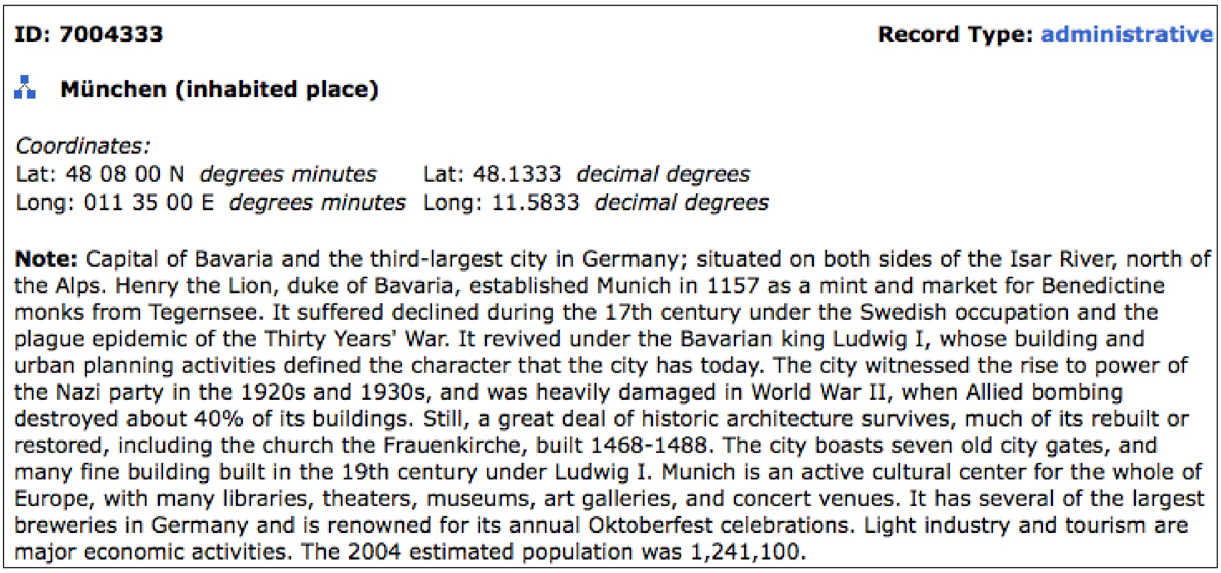
- comprises a hierarchy with tree structures corresponding to the current and historical worlds;

- use of current terminology, warranted for use by authoritative literary sources, and validated by use in the scholarly art and architectural history community;

- compiled and edited in response to the needs of the user community.

Currently, TGN contains around 1,106,000 names and other information about places. Names for a place may include names in the vernacular language, English, other languages, historical names, names in natural and inverted order. Among these names, one is flagged as the preferred name, that is, the main designation when referring to a place.

Its structure consists of a hierarchical database, where the focus of each record is a place. Each place record is identified by a unique numeric ID. Linked to the record for the place are additional (and some optional) fields: names, the place's parent or position in the hierarchy, other relationships, geographic coordinates, notes, sources for the data, and place types, which are terms describing the role of the place (e.g., inhabited place and state capital). The temporal coverage of the TGN ranges from prehistory to the present and the scope is global. Figure 2.2 shows part of a TGN record regarding the german city of München, where we can view its system's identifier (ID), name, type, coordinates and description.



ID: 7004333 **Record Type:** administrative

München (inhabited place)

Coordinates:
Lat: 48 08 00 N *degrees minutes* Lat: 48.1333 *decimal degrees*
Long: 011 35 00 E *degrees minutes* Long: 11.5833 *decimal degrees*

Note: Capital of Bavaria and the third-largest city in Germany; situated on both sides of the Isar River, north of the Alps. Henry the Lion, duke of Bavaria, established Munich in 1157 as a mint and market for Benedictine monks from Tegernsee. It suffered declined during the 17th century under the Swedish occupation and the plague epidemic of the Thirty Years' War. It revived under the Bavarian king Ludwig I, whose building and urban planning activities defined the character that the city has today. The city witnessed the rise to power of the Nazi party in the 1920s and 1930s, and was heavily damaged in World War II, when Allied bombing destroyed about 40% of its buildings. Still, a great deal of historic architecture survives, much of its rebuilt or restored, including the church the Frauenkirche, built 1468-1488. The city boasts seven old city gates, and many fine building built in the 19th century under Ludwig I. Munich is an active cultural center for the whole of Europe, with many libraries, theaters, museums, art galleries, and concert venues. It has several of the largest breweries in Germany and is renowned for its annual Oktoberfest celebrations. Light industry and tourism are major economic activities. The 2004 estimated population was 1,241,100.

Figure 2.2: Excerpt of a record from the TGN

The trees in the hierarchical database descend from a top root from which there can be more than one hierarchy grown. Currently most of the terms in the TGN are organized under a facet ”World” where its constituent hierarchies represent the current political and physical world, although some historical nations and empires are also included.

In addition to the hierarchical relationships, the TGN has equivalent and associative relationships, making it compliant with the standards for being a thesaurus[4]:

[4] For a more detailed description and definition of the concept of thesaurus please see the standards ISO 2788 (for monolingual thesauri) and ISO 5964 (for multilingual thesauri)

Definition 4. *a **thesaurus** is a semantic network of unique concepts, including relationships between synonyms, broader and narrower (parent/child) contexts, and other related concepts. Thesauri allow three types of relationships: equivalence (synonym), hierarchical (whole/part or generic/species), and associative.*

Therefore the TGN is considered a geographic thesaurus. While many of its records include coordinates, these coordinates are approximate and are intended for reference (”finding purposes”) only. It can not be considered an accurate source for georeferencing and also does not supply coordinates for every place.

The project is of historic value, because it was one of the first efforts in the area of digital geographic information and serves as a resource for both research and content purposes.

2.3.2 ADL-G - Alexandria Digital Library Gazetteer

The Alexandria Digital Library (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 academic, public, and private 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 [34].

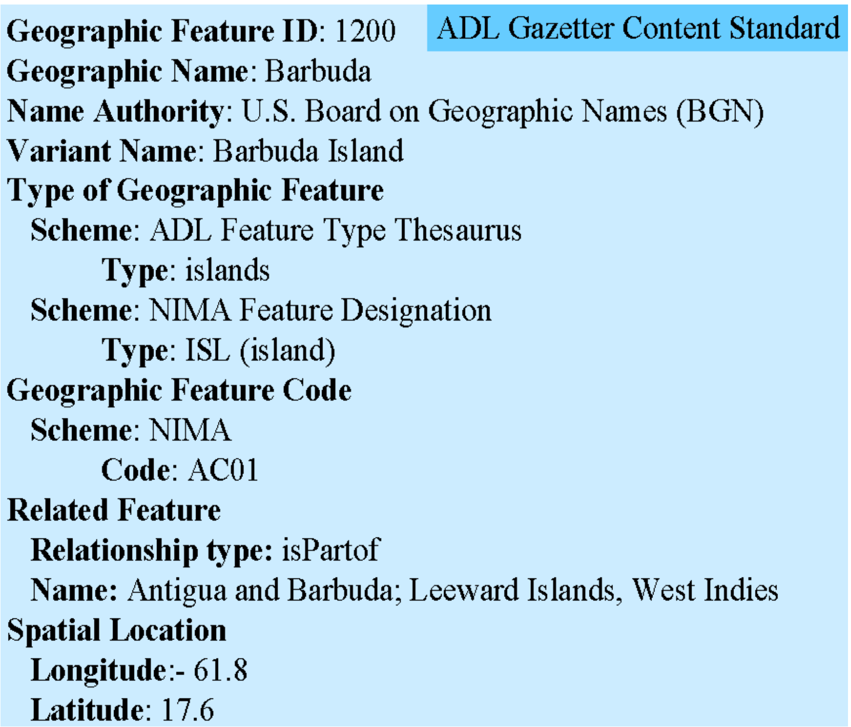
The project required a gazetteer in order to provide their spatial query functions so as to georeference the objects in the library. This was accomplished with the merging of databases acquired from two U.S. federal agencies - U.S. National Imagery and Mapping Agency (NIMA) with their *Geographic Names Processing System* (GNPS) and the U.S. Geological Survey with their *Geographic Names Information System* (GNIS) plus U.S. topographic map areas, California earthquake epicenters, volcanoes from the Smithsonian Global Volcanism Program, and additional sets of bounding boxes for administrative areas.

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 so there could be mappings between various type schemas [32].

The solution for these issues was the development of the ADL Gazetteer Content Standard (GCS) and the ADL 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 [34]. Figure 2.3 [34] shows an example of the description of an element in the GCS. In it, we can highlight the Feature ID, name, feature type (both from the ADL FTT and NIMA), the relationships with other features and the spatial footprint.



Geographic Feature ID: 1200 ADL Gazetter Content Standard

Geographic Name: Barbuda

Name Authority: U.S. Board on Geographic Names (BGN)

Variant Name: Barbuda Island

Type of Geographic Feature

Scheme: ADL Feature Type Thesaurus

Type: islands

Scheme: NIMA Feature Designation

Type: ISL (island)

Geographic Feature Code

Scheme: NIMA

Code: AC01

Related Feature

Relationship type: isPartof

Name: Antigua and Barbuda; Leeward Islands, West Indies

Spatial Location

Longitude:- 61.8

Latitude: 17.6

Figure 2.3: GCS example

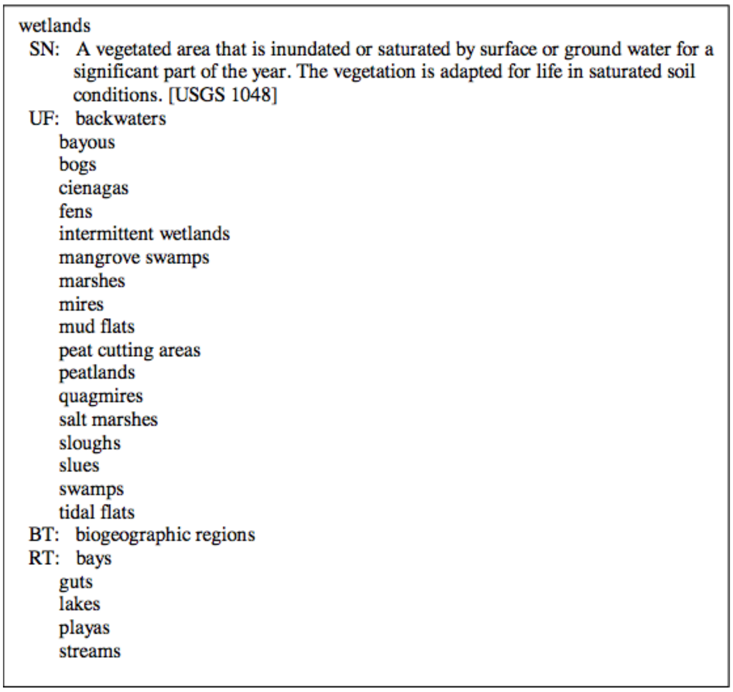
The purpose of the FTT was to try and establish a common link between the various typing schemas by adding, to the greatest extent possible, all of the vocabulary of the other schemas either as preferred terms, based on their average use in reference sources or dictionaries, or as alternate names pointing 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. In Figure 2.4 [34], we can see an example of a term entry in the FTT. In this example we have the entry for *wetlands* which shows its definition (SN - Scope Note), the alternative terms which can indicate this entry (UF - Used For), the category of terms where this entry relates (BT - Broader terms) and other entries which are also preferred terms and belong in the same category as this entry (RT - Related Terms).

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

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



wetlands
SN: A vegetated area that is inundated or saturated by surface or ground water for a significant part of the year. The vegetation is adapted for life in saturated soil conditions. [USGS 1048]
UF: backwaters
bayous
bogs
cienagas
fens
intermittent wetlands
mangrove swamps
marshes
mires
mud flats
peat cutting areas
peatlands
quagmires
salt marshes
sloughs
slues
swamps
tidal flats
BT: biogeographic regions
RT: bays
guts
lakes
playas
streams

Figure 2.4: FTT entry example

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;

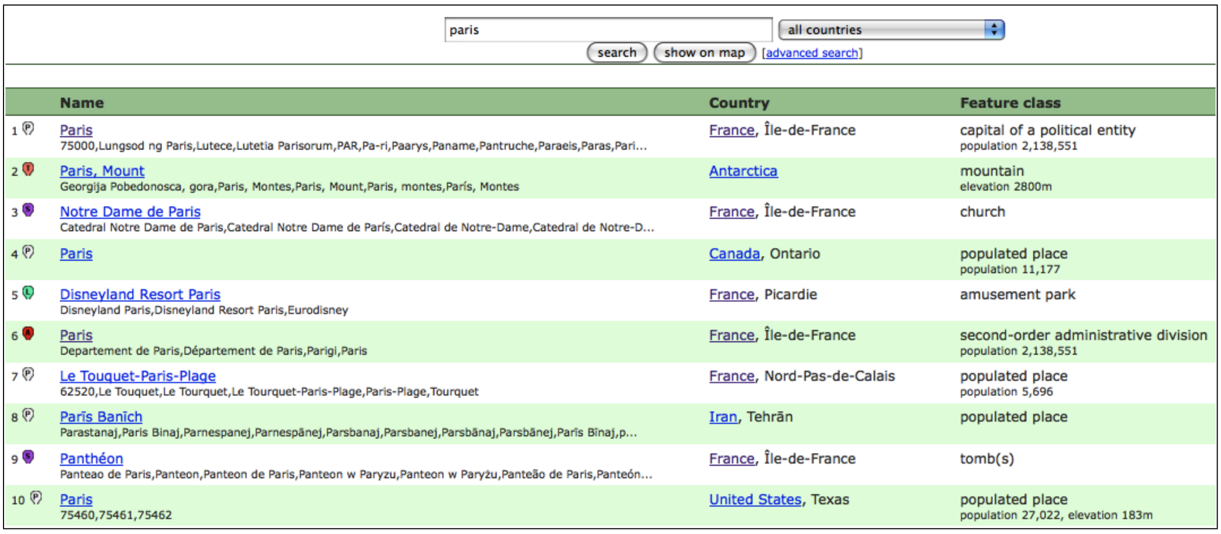
The reports may be of two types: normal or extended. The main difference between them is the amount of information they contain. The normal one contains only some elements of a gazetteer entry, namely its system identifier, names, footprints, relationships and type classifications. The extended report 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.

2.3.3 Geonames

The Geonames project was founded by swiss software engineer Marc Wick and was 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.

It contains over 8 million geographical names and consists of 7 million unique features whereof 2.6 million populated places and 2.8 million alternate names in up to 200 languages. This information is retrieved from various sources (over 100) but most importantly from the National Geospatial-Intelligence Agency's (NGA) and the U.S. Board on Geographic Names, GNIS, GeoBase[6]. Users are able to manually edit, correct and add new names via a user-friendly wiki interface and the data is accessible free under a creative commons attribution license[7], through a number of web-services and a daily database export. Figure 2.5 shows the top 10 feature results in a search for ”Paris”.



	Name	Country	Feature class
1	Paris 75000,Lungsod ng Paris,Lutece,Lutetia Parisorum,PAR,Pa-ri,Paarys,Paname,Pantruche,Paraeis,Paras,Pari...	France, Île-de-France	capital of a political entity population 2,138,551
2	Paris, Mount Georgija Pobedonosca, gora,Paris, Montes,Paris, Mount,Paris, montes,París, Montes	Antarctica	mountain elevation 2800m
3	Notre Dame de Paris Catedral Notre Dame de Paris,Catedral Notre Dame de París,Catedral de Notre-Dame,Catedral de Notre-D...	France, Île-de-France	church
4	Paris	Canada, Ontario	populated place population 11,177
5	Disneyland Resort Paris Disneyland Paris,Disneyland Resort Paris,Eurodisney	France, Picardie	amusement park
6	Paris Departement de Paris,Département de Paris,Parigi,Paris	France, Île-de-France	second-order administrative division population 2,138,551
7	Le Touquet-Paris-Plage 62520,Le Touquet,Le Tourquet,Le Tourquet-Paris-Plage,Paris-Plage,Tourquet	France, Nord-Pas-de-Calais	populated place population 5,696
8	Parīs Banīch Parastanaj,Paris Binaj,Parnespanej,Parnespānej,Parsbanaj,Parsbanej,Parsbānaj,Parsbānej,Parīs Bīnaj,p...	Iran, Tehrān	populated place
9	Panthéon Panteao de Paris,Panteon,Panteon de Paris,Panteon w Paryzu,Panteon w Paryżu,Panteão de Paris,Panteón...	France, Île-de-France	tomb(s)
10	Paris 75460,75461,75462	United States, Texas	populated place population 27,022, elevation 183m

Figure 2.5: Screenshot of a Geonames search list

The motivation for the project came when Wick was developing a holiday apartments application and required gazetteer data. Due to cost restrictions on commercial data, he began searching and collecting free data from various sources on the Web and subsequently releasing it as the Geonames project. The idea was to share the data he had acquired and at the same time improve its quality and size, seeing as that a collective effort at building a gazetteer would be far more successful than several individual efforts.

A lot of applications shared this point of view and subsequently switched from their own proprietary solutions to Geonames. Among them are Greenpeace, the British Broadcasting Company (BBC) and LinkedIn[8]. Also, Wikipedia has several entries which are ”geotagged” to entries in Geonames in order to complement its database.

With several members of the Geonames team serving as *ambassadors*, who help with questions regarding their countries and serve as a local contact person for national data providers such as national mapping agencies, statistical offices or postal services, 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 [37] of geographic features:

Definition 5. *an* ***ontology*** *is an explicit specification of a concept used to achieve a shared and common understanding of a particular domain of interest.*

Meaning we 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) [37].

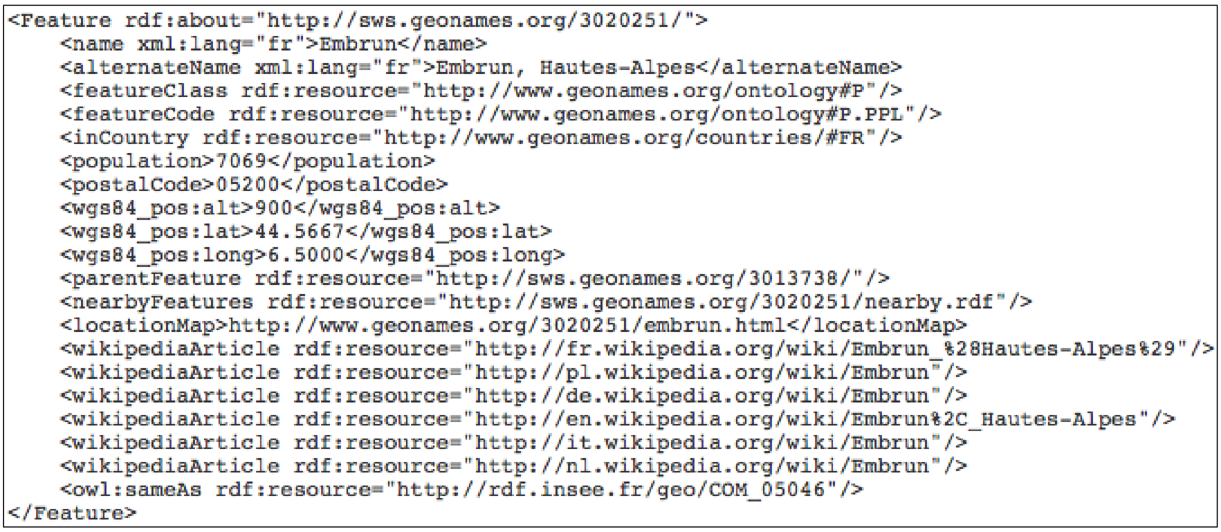
In Geonames, this is done by means of the Web Ontology Language (OWL). OWL is a knowledge representation language for producing ontologies based on Resource Description Framework (RDF). RDF

[6] `http://www.geobase.ca`
[7] `http://en.wikipedia.org/wiki/Creative_Commons_licenses`
[8] `http://www.linkedin.com`

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. The application of the ontology to Geonames is made by the provision of two URL for each place name: one referring to the place name itself and another to an OWL document describing it. The way Geonames describes it, its web service is using two URL to clearly distinguish between Concept (the entity as is) and Document (the document with the information pertaining the entity). An example is the following URL for the french town Embrun (the associated content is shown in the Figure 2.6):

1. `http://sws.geonames.org/3020251/`

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



```
<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 2.6: Example of a Geonames feature's description coded in 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 (neighboring countries)

- adjacent features

An example of RDF description of a Geonames feature is in Figure 2.6, 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`.

The service is currently perhaps the most widely accepted geographic resource and also one of the most used geographic databases if not the most used.

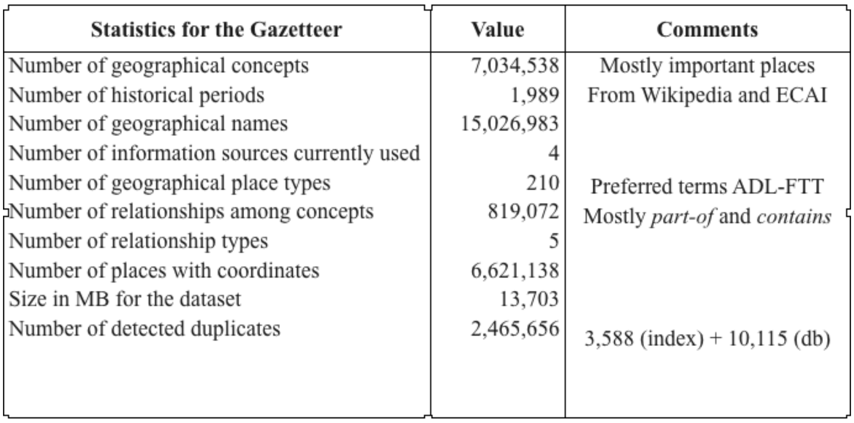
2.3.4 DIGMAP

The Discovering our Past World with Digitized Maps (DIGMAP) project [11] 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 on-line 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[9] 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. Figure 2.7 shows a table showing a statistical characterization of the data gathered in the gazetteer.



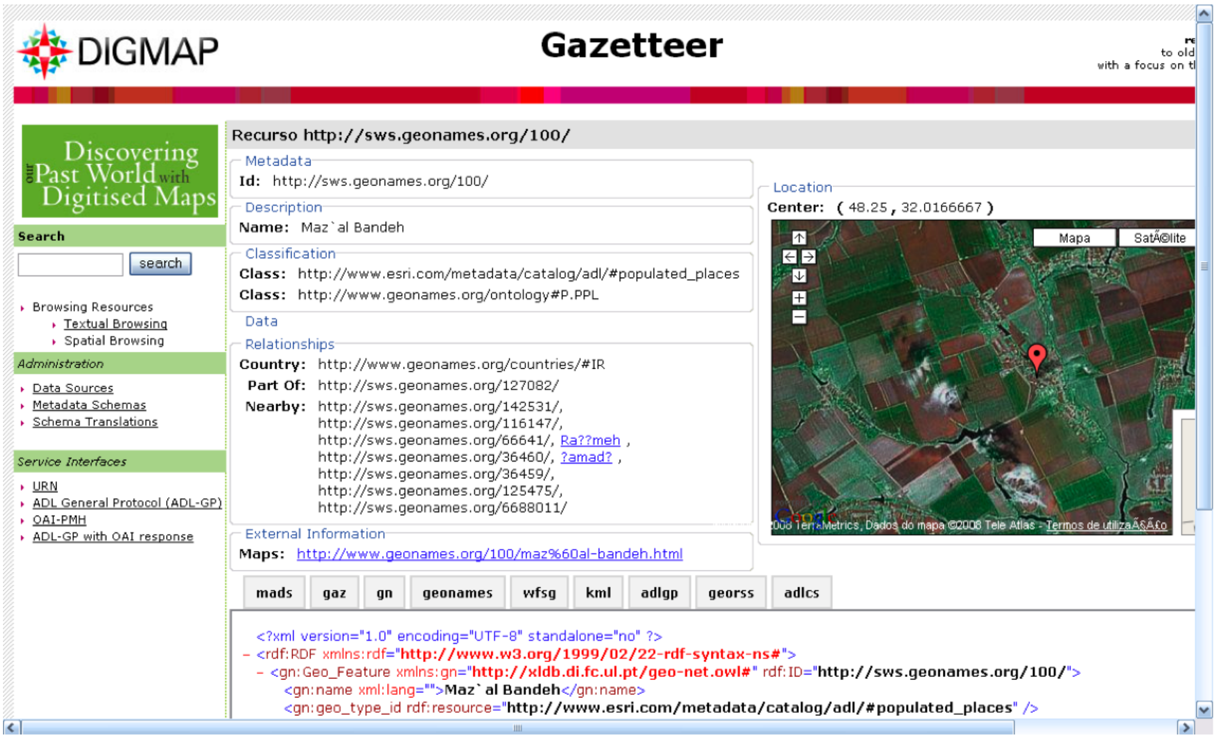
Statistics for the Gazetteer	Value	Comments
Number of geographical concepts	7,034,538	Mostly important places
Number of historical periods	1,989	From Wikipedia and ECAI
Number of geographical names	15,026,983	
Number of information sources currently used	4	
Number of geographical place types	210	Preferred terms ADL-FTT
Number of relationships among concepts	819,072	Mostly *part-of* and *contains*
Number of relationship types	5	
Number of places with coordinates	6,621,138	
Size in MB for the dataset	13,703	
Number of detected duplicates	2,465,656	3,588 (index) + 10,115 (db)

Figure 2.7: Statistical categorization of the DIGMAP gazetteer data

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.). An example of a feature from the gazetteer displayed on the portal can be seen in figure 2.8.

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

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



Recurso http://sws.geonames.org/100/

Metadata
Id: http://sws.geonames.org/100/

Description
Name: Maz`al Bandeh

Classification
Class: http://www.esri.com/metadata/catalog/adl/#populated_places
Class: http://www.geonames.org/ontology#P.PPL

Data

Relationships
Country: http://www.geonames.org/countries/#IR
Part Of: http://sws.geonames.org/127082/
Nearby: http://sws.geonames.org/142531/,
http://sws.geonames.org/116147/,
http://sws.geonames.org/66641/, Ra??meh ,
http://sws.geonames.org/36460/, ?amad? ,
http://sws.geonames.org/36459/,
http://sws.geonames.org/125475/,
http://sws.geonames.org/6688011/

Location
Center: (48.25 , 32.0166667)

2008 TerraMetrics, Dados do mapa ©2008 Tele Atlas - Termos de utilizaÃ§Ã£o

External Information
Maps: http://www.geonames.org/100/maz%60al-bandeh.html

```
  <?xml version="1.0" encoding="UTF-8" standalone="no" ?>
- <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
  - <gn:Geo_Feature xmlns:gn="http://xldb.di.fc.ul.pt/geo-net.owl#" rdf:ID="http://sws.geonames.org/100/">
      <gn:name xml:lang="">Maz`al Bandeh</gn:name>
      <gn:geo_type_id rdf:resource="http://www.esri.com/metadata/catalog/adl/#populated_places" />
```

Figure 2.8: DIGMAP gazetteer feature

2.4 Linked Data

With the rise of the World Wide Web, there was a dramatic change in the way people share information by presenting them with a means to publish and access documents in a global space of information. By combining individual Hypertext Markup Language (HTML) documents with Hyperlinks, we obtained a Web of documents. These documents were then traversed through the use of Web browsers and later on, indexed and analyzed by search engines so as to provide a means to infer structure and relevance between them so as to better provide response for user queries. All of this made possible through the simple and extensible nature of the Web, which in turn, enabled its tremendous growth. However, with the passing of the years, the Web became not only a space for linked documents but one where both documents and data are published. But the same principles that allow this global space of linked documents to become so vast weren't effective for data, mainly because of HTML's expressive limits.

2.4.1 The Concept

Recently, the Web of documents has become also a Web of data. People started weaving individual bits of data together with RDF triples that expressed the relationship between these bits of data, forming this Web of data. This is because of the publication of a set of best practices [9] by Tim Berners-Lee for publishing and linking structured data on the World Wide Web in 2006. They are:

- Use Uniform Resource Identifier (URI)s to identify things.
- Use HTTP URIs so that these things can be referred to and looked up by people and user agents.
- Provide useful information about the thing when its URI is dereferenced, using standard formats such as RDF or XML.
- Include links to other, related URIs in the exposed data to improve discovery of other related information on the Web.

In common usage, Linked Data refers to these principles. These principles stem from the author's vision to conceive a Semantic Web by putting or converting data to be published in a form that machines can naturally understand, thus creating a Web of data that can be processed directly or indirectly by machines, their meaning is explicitly defined, it is linked to other external data sets, and can in turn be linked to from external data sets [8].

Linked Data is no more complex than this - connecting related data across the Web relying on three important technologies: URI, HTTP and RDF. The RDF model encodes data in the form of subject, predicate, object triples. The subject and object of a triple are both URIs that each identify a resource, or a URI and a string literal respectively. The predicate specifies how the subject and object are related, and is also represented by a URI.

This way, one can relate two people (that are described by RDF links), relate a movie to a genre, a song to an artist, etc; basically, we can link any real-world entities through their RDF descriptions and the use of vocabularies (collections of classes and properties which serve to describe entities and how they are related). These vocabularies themselves are described in RDF and as such, anyone can create a vocabulary and publish it in order to create new relations and mappings between already existing relations. And, by using URI's to identify resources and HTTP to retrieve them, LInked Data builds upon the existing Web architecture in order to create something new, adding new properties and giving it more power of expression.

This has lead to an extension of the World Wide Web to include a global data space from diverse areas such as cinema, music, television, radio, books, scientific and statistical data, etc. And, as with the rise of the Web of documents, the appearance of new types of applications to browse this global data space.

Like the regular Web browsers follow Hypertext links to travel between HTML documents, Linked Data browsers follow links of RDF triples to navigate between data. Some may be direct applications of the normal hyperlink traversal approach, such as the Disco[10] browser which simply navigates between RDF links of one data set to another. However, other browsers such as Marbles[11] try to take advantage of the vast amount of data by gathering and merging information about the same entity from different data sources in order to provide a complete description of an entity, while providing data provenance. Still, there is the limitation of not being able to navigate by keywords, being limited to knowing what URI you want to go to or navigating through the links in hopes of reaching the intended entity. This is where the search engines come in.

In the case of search engines, the Linked Data has two types of approaches: human-oriented engines, which are the traditional approach and application-oriented engines. The human-oriented ones follow the traditional interaction set by traditional search engines such as Google. A user can type keywords into a text box and the application returns a set of results most closely related to the keywords provided. An example is Falcons[12] , a search engine that traverses through embedded RDF information on web pages and produces a list of results, depending on the type of entity selected. Also, it provides a short description of each entity in the result list for better choosing.

The other approach relates to a need to provide means for Linked Data applications not have to query the entire Web of Data to search for related entities. These indexes provide means for Linked Data applications to discover RDF documents on the Web that reference a certain URI or contain certain keywords. An example is the Sindice [13] index, which serves as both a semantic web search engine and semantic aggregation service.

[10] http://www4.wiwiss.fu-berlin.de/bizer/ng4j/disco/

[11] http://wiki.dbpedia.org/Marbles?v=71e

[12] http://iws.seu.edu.cn/services/falcons/

[13] http://sindice.com

Besides these types of applications, there has also been the development of more domain-targeted applications. These applications act as “mash-ups” of data from Linked Data sources in order to provide more domain-specific functionality. According to a technical report made in 2009 [28], these applications can be roughly divided into four categories, from a Linked Data usage point-of-view:

- *content reuse*: applications that mainly reuse content of datasets in the LOD cloud in order to safe time and resources;
- *semantic tagging and rating*: applications that use URIs in the datasets for unambiguously talking about things;
- *integrated question-answering*: applications that focus on answering a user’s question;
- *event data management systems*: applications that allow people to organize and query event-related data.

In the case of content reuse, an example is the BBC Music site[14] build around the Musicbrainz[15] and DBpedia metadata.

Revyu[16] is a generic reviewing and rating site based on Linked Data principles. In addition to publishing linked data, this Web application consumes linked data to enhance the end-user’s experience, exploiting the interlinking with DBpedia.

DBpedia Mobile[17] is an interesting application for mobile environments; basically it is a location-centric DBpedia client application for mobile devices, that is - based on the GPS signal of a mobile - able to render a map indicating nearby locations from the DBpedia dataset. It serves to roughly answer the question “What interesting places are around me?”.

2.4.2 Linked Open Data

Besides the concept of Linked Data, there is another important name that often generates confusion by using it to refer to the concept as well. While Linked Data refers to the concept of the Web of Data, its principles and characteristics, Linking Open Data (LOD)[18] refers to an community initiative or project that relates to the *Open Data Movement* - a philosophy or practice that aims at making data freely available to anyone. This is done by interconnecting several open data sets already available on the Web, such as Wikipedia, Geonames, among many others which are published under Creative Commons[19] licenses.

The goal of the Linking Open Data community project is to extend the Web with a data commons by publishing various open data sets as RDF on the Web and by setting RDF links between data items from different data sources. Also, to show the importance and value of the Semantic Web, it is essential to have more real-world data online and this project incentives the publication of more and different types of data for inter-linking with the already available ones in order to create a global data network, equally accessible by both man and machine.

Figure 2.9 shows the data sets that have been published and interlinked by the project so far. Collectively, these 203 data sets consist of over 25 billion RDF triples, which are interlinked by around 395 million RDF links (these values stand as of September 2010).

[14] `http://www.bbc.co.uk/music/`

[15] `http://musicbrainz.org/`

[16] `http://revyu.com/`

[17] `http://beckr.org/DBpediaMobile/`

[18] `http://esw.w3.org/SweoIG/TaskForces/CommunityProjects/LinkingOpenData`

[19] `http://creativecommons.org/`

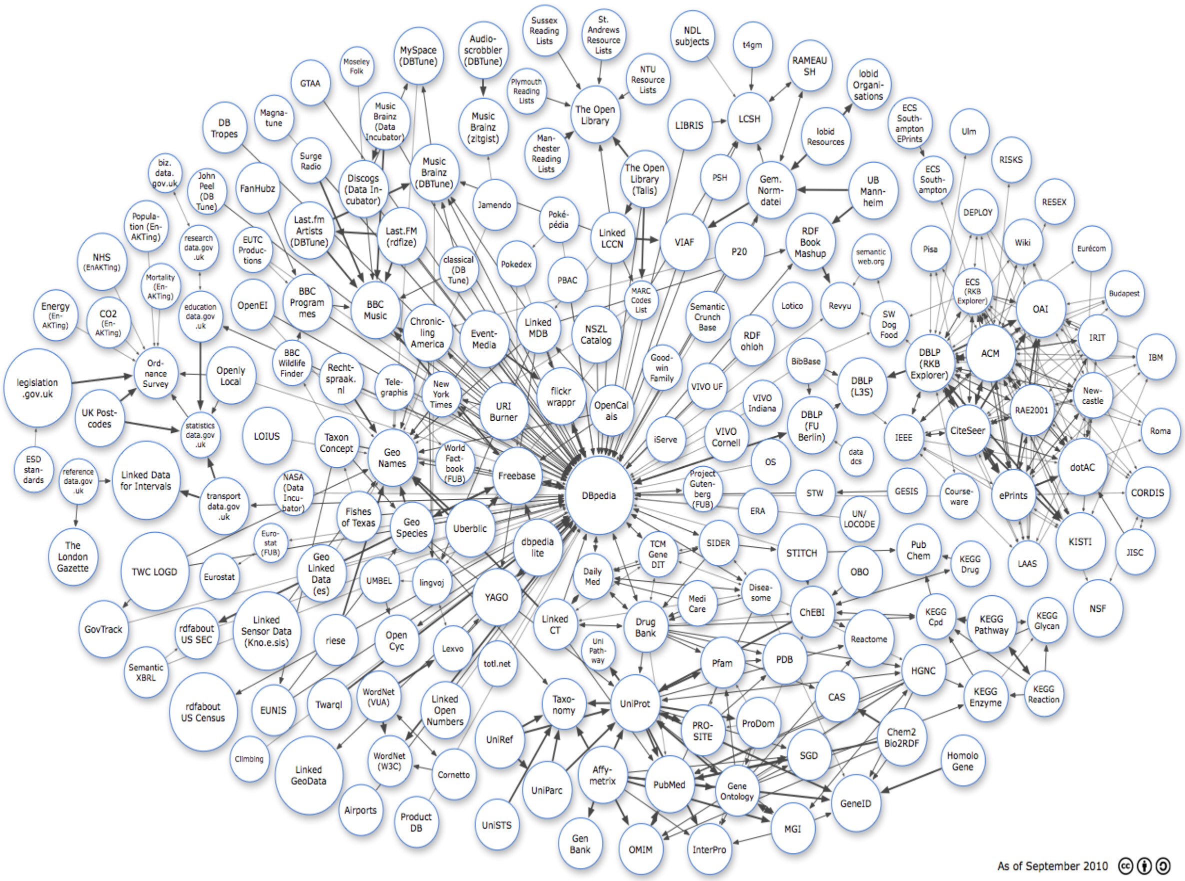


Figure 2.9: Current LOD cloud diagram

2.5 SRU protocol

Search/Retrieve URL Service (SRU)[20] is a Web Service-based protocol for querying Internet indexes or databases and returning search results. Its goal is to define a standard form for Internet search queries as well as the structure of the responses. Much of the functionality of SRU is derived from its antecedent protocol, the Z39.50[21], however, in a simplified form. Specifications for the protocol were first published in 2002 and have been popular for use in new applications because of its ease of implementation.

SRU is a REST-ful Web service. The end user creates a search request on the user system, which employs a specific local query syntax. In this protocol's case, the syntax used is Contextual Query Language (CQL), a formal language for representing queries to information retrieval systems such as web indexes, bibliographic catalogs and museum collection information. The objective is that queries be human readable and writable, and that the language be intuitive while maintaining the expressiveness of more complex languages. The server then reads the input, processes it, and returns the results as an XML stream back to the client. REST-ful Web Services encode the query input usually in the shape of URLs. Each name/value pair of the query string specifies a set of input parameters for the server. Once received, the server parses these name/value pairs, does some processing using them as input, and returns the results as an XML stream. The shape of the query string as well as the shape of the XML stream are dictated by the protocol.

[20] http://www.loc.gov/standards/sru/
[21] http://www.loc.gov/z3950/agency/

2.5.1 Operations

SRU has three main operations: *Explain* which presents the SRU service's description and provided functionality, *Scan* which presents listings of matching terms in the database for a particular query and *Search/Retrieve* which fetches lists of database records for a query. Contrary to the Scan Operation which returns only the identifier or name of the matching record in the database, the Search/Retrieve operation retrieves the full database records.

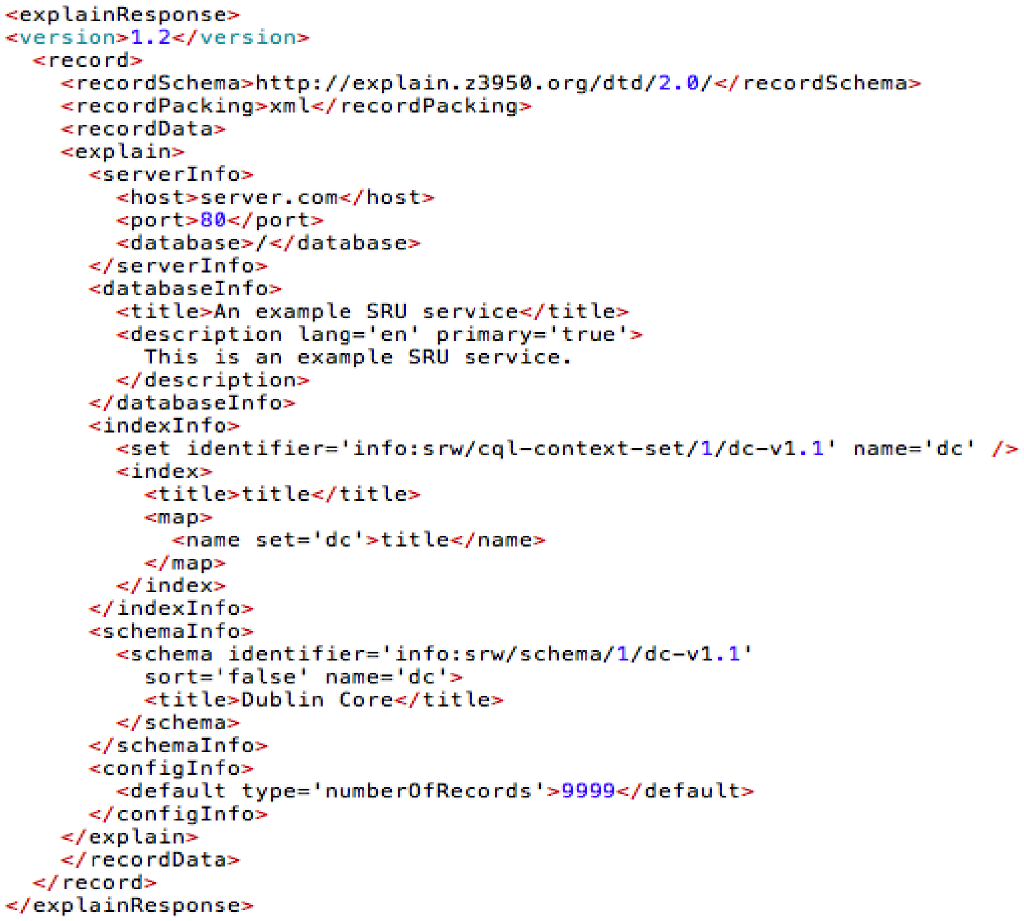
Explain

The Explain operation is a request sent by clients as a way of learning about the server's database/index as well as its functionality. At a minimum, responses to explain operations return the location of the database, a description of what the database contains, and what features of the protocol the server supports. Implemented in SRU, empty query strings on a URL are interpreted as an explain operation. It can also be explicitly invoked, in which case, a version parameter is mandatory. So, the explain record can be obtained in two ways:

- `http://server.com/`
- `http://server.com/?operation=explain&version=1.2`

One example of an SRU explain record can be like in the case of Figure 2.10 below. From its content, one can ascertain the following:

- the server supports version 1.2 of the protocol
- records in this response are encoded in a specific DTD and are encoded as XML
- the location of the server
- a description of the database
- the database search indexes supported (in this case, only search by title)
- the database return schema used is Dublin Core
- a maximum number of records returned (no more than 9999)



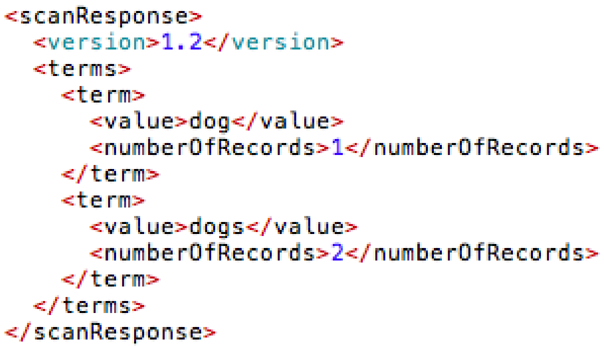
```
<explainResponse>
<version>1.2</version>
  <record>
    <recordSchema>http://explain.z3950.org/dtd/2.0/</recordSchema>
    <recordPacking>xml</recordPacking>
    <recordData>
    <explain>
      <serverInfo>
        <host>server.com</host>
        <port>80</port>
        <database>/</database>
      </serverInfo>
      <databaseInfo>
        <title>An example SRU service</title>
        <description lang='en' primary='true'>
          This is an example SRU service.
        </description>
      </databaseInfo>
      <indexInfo>
        <set identifier='info:srw/cql-context-set/1/dc-v1.1' name='dc' />
        <index>
          <title>title</title>
          <map>
            <name set='dc'>title</name>
          </map>
        </index>
      </indexInfo>
      <schemaInfo>
        <schema identifier='info:srw/schema/1/dc-v1.1'
          sort='false' name='dc'>
          <title>Dublin Core</title>
        </schema>
      </schemaInfo>
      <configInfo>
        <default type='numberOfRecords'>9999</default>
      </configInfo>
    </explain>
    </recordData>
  </record>
</explainResponse>
```

Figure 2.10: SRU Explain response example

Scan

Scan operations list and enumerate the terms found in the remote database's index. Clients send scan requests and servers return lists of terms. This enables clients to present an ordered list of values and, if supported, how many hits there would be for a search on that term. Scan is often used to select terms for subsequent searching or to verify a negative search result. Here is a basic request and response:

- `http://server.com/?operation=scan&scanClause=dog&version=1.2`



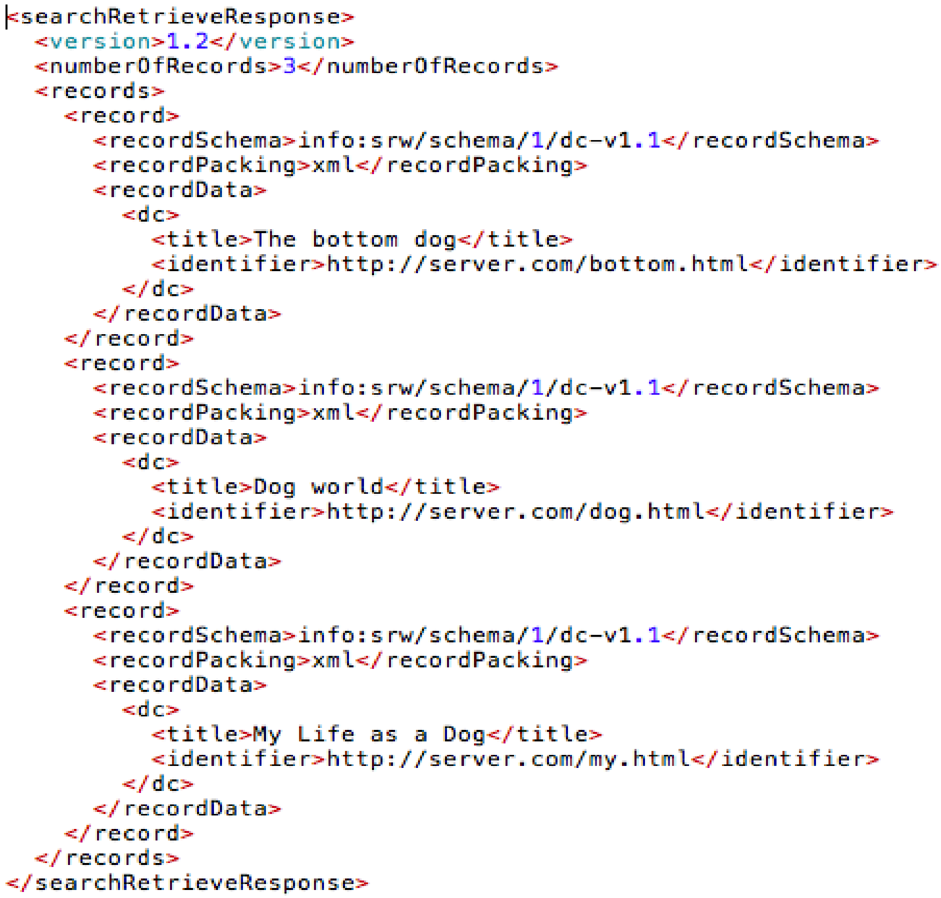
```
<scanResponse>
  <version>1.2</version>
  <terms>
    <term>
      <value>dog</value>
      <numberOfRecords>1</numberOfRecords>
    </term>
    <term>
      <value>dogs</value>
      <numberOfRecords>2</numberOfRecords>
    </term>
  </terms>
</scanResponse>
```

Figure 2.11: SRU Scan response example

Search/Retrieve

SearchRetrieve operations are the main operation in SRU. They provide the means to query the remote database and return search results. Queries must be articulated using the Contextual Query Language (CQL). Servers do not have to implement every aspect of the CQL syntax, but they have to know how to return diagnostic messages when something is requested but not supported. The results of searchRetrieve operations can be returned in any number of formats, as specified via explain operations. Below is a simple request for documents matching the free text query 'dog':

- `http://server.com/?operation=searchRetrieve&query=dog&version=1.2`



```
<searchRetrieveResponse>
  <version>1.2</version>
  <numberOfRecords>3</numberOfRecords>
  <records>
    <record>
      <recordSchema>info:srw/schema/1/dc-v1.1</recordSchema>
      <recordPacking>xml</recordPacking>
      <recordData>
        <dc>
          <title>The bottom dog</title>
          <identifier>http://server.com/bottom.html</identifier>
        </dc>
      </recordData>
    </record>
    <record>
      <recordSchema>info:srw/schema/1/dc-v1.1</recordSchema>
      <recordPacking>xml</recordPacking>
      <recordData>
        <dc>
          <title>Dog world</title>
          <identifier>http://server.com/dog.html</identifier>
        </dc>
      </recordData>
    </record>
    <record>
      <recordSchema>info:srw/schema/1/dc-v1.1</recordSchema>
      <recordPacking>xml</recordPacking>
      <recordData>
        <dc>
          <title>My Life as a Dog</title>
          <identifier>http://server.com/my.html</identifier>
        </dc>
      </recordData>
    </record>
  </records>
</searchRetrieveResponse>
```

Figure 2.12: SRU searchRetrieve response example

In this case, the server returns three hits and by default includes Dublin Core title and identifier elements.

Chapter 3

Technical and Research Challenges in Gazetteer Development

This chapter presents some of the more active areas of concern currently in gazetteer development. These include research of a different form of gazetteers (event gazetteers), Volunteered Geographic Information, techniques and practices in data quality control and data integration from different sources and, more specifically, a survey of the current state of the art in duplicate detection techniques and adapting the problem towards the context of gazetteer record linkage.

3.1 Event Gazetteers: spatio-temporal referencing

Gazetteers have become increasingly important tools for place-name reference and querying and have also been assigned new roles of importance in modern information systems [33]. These all relate to spatial referencing capabilities, however one must also consider the temporal side of the spectrum. Since no place in the real world is eternal or can have attributes that are persistent, one can consider every named place as an event [49].

Definition 6. *An* ***event*** *is considered as an instance of information in the form of a named, temporally and spatially delimited entity.*

A gazetteer directed to events can have several applications but the most notable ones would be for historical and cultural purposes. With an historical event gazetteer, one can reference people, places or even abstract notions on events located in time and space while having a sense of the broader context in which those events occurred. Examples of that include browsing through descriptions about the evolution of species during the prehistoric eras, comparing the different historic name attributions to the chinese city of Beijing thru out its existence or viewing descriptions of all the battles and invasions occurred during the Napoleonic Wars, among other things.

There have been several efforts in order to model event descriptions such as the History Events Mark Up and Linking (HEML) project to develop Extensible Markup Language (XML) schemas for historic events, though dealing with them as isolated cases [53].

Another type of project is the Event Structure Analysis (ESA) [29] created in the 1980's and developed over the past 20 years. ESA creates sequences of social events that are linked both chronologically and causally (signifying that prior events are responsible for the appearance of new events) providing historically based causal interpretations of history. It does this by analyzing source material in order to create discrete events, pondering about logical chronologies where the events fit, creating logical relations between the events and establishing causality by stipulating some events are prerequisites for other ones.

Other researchers also use this focus on event relationships in order to expose different perpectives for the same historic event by creating parallel timelines depicting different points of view of people involved or related to an event and also make it possible to associate which elements of a complex event were visible to its different participants.

Examples of this type of work are the Temporal Modeling Project (TMP) [17] and SemTime [38]. Another example is the work in [6] where they presented a gazetteer that stored events with attributes such as type, location, beginning and end times and actors, grouping these events into larger sets which constituted events themselves, but lacked work in explaining causality. However these projects dealt mostly with temporal aspects, that is, they did not concern themselves with mapping the event to real world locations.

More recent research efforts have attempted to create practices and methodologies into incorporating both spatial and temporal aspects of visualizing events. The DIGMAP project considers the ever changing aspect of geography across time and so incorporates a temporal domain to its gazetteer so as to increase the potential of querying in their system for regions that have changed names, merged, changed borders among other possibilities [44].

In [48] the author proposes the establishing of design principles for gazetteers so they can be better suited for use by historians and humanists following with [49] where she presents the Rethinking Timelines project, which builds on the TimeMap Project[1] where they attempt to create compelling visual representations of spatio-temporal entities, while discussing the concept of historical gazetteers and its desirable implementation as a rich and reliable source for historical data both temporally and spatially located.

3.2 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 [33], 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 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

[1] http://www.timemap.net

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. In the process of filling the ADL gazetteer, one of their main concerns was the conversion of the database data they imported to their own schema and dealing with the subsequent problems of differences in typing schemas; they grouped them mainly as problems of: specificity (where the incoming data has more specific terms than their schema); generality (where the incoming data has broader categories than their schema); definition (where it isn't clear how to interpret the scope of an incoming category); and no category (where there has been no attempt to categorize the placenames) [32].

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 [12, 13, 60]. 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 problem is the process of merging similar geospatial entities, which is commonly referred to as feature conflation. It drives from the existence of different feature representations referring to the same physical object/region in space and combine them into features that possess greater knowledge and accuracy than the original sources.

The essence is that all attributes of gazetteer entry may be imprecise, because the features to which they refer tend to be approximate in space and can change with time, so the same features tend to acquire multiple feature names and sometimes types.

Work in the ADL has been done to deal with existence of "duplicates" in their gazetteer [26] and there has been research to try and achieve gazetteer conflation [27] and more recently with the use of semantic processes [52]. Since this is a pertinent requirement for this dissertation work, it will be discussed in more detail in its own section further ahead.

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 [50]; 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 [59].

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 [5]; 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) [47] and the use of mining techniques for information concerning known place names in the vague region from the Web [40].

One other form of enriching features is the incorporation of shapefiles[2] as another attribute. Shapefiles are a geospatial vector data format storing geometric location and associated attribute information. It

[2] For a complete technical description, visit `http://www.esri.com/library/whitepapers/pdfs/shapefile.pdf`

was developed by ESRI[3]. Shapefiles describe geometries such as points, lines, and polygons. These, in turn, could represent features like water wells, rivers, and lakes, respectively. Each item may also have attributes that describe them, such as the name or temperature. From their description, we can infer their usefulness as a form of mapping visual representation for features in the gazetteer.

3.3 VGI - Benefits and concerns

The rise and expansion of social computing and networking practices has served to substantially increase the amount of user-generated content online. In the context of the Web, particularly the Web 2.0, user-generated content refers to the ability of Web users to create content that is then integrated and made available through Web sites. This has had a significant impact on geographic information over the past couple of years: there has been a substantial increase of interest of the general public in the creation, use and distribution of geographic information provided by them. This information is provided mostly voluntarily and since the providers are usually people with low or no qualifications in the area, the information may or may not be accurate.

The term created to define this collective movement was that of Volunteered Geographic Information (VGI) [21] and has become popular thanks to sites like Wikimapia. An online map and satellite imaging resource whose aim is ”to describe the whole world”, it combines Google Maps[4] with a wiki system and currently offers over 11,5 million places that have been identified and annotated by its users. It’s a clear example of web sites established for the purpose of inviting and assembling data about places in a free nature as is OpenStreetMap[5] which manages to outdo commercial mapping sites in terms of update frequency and thematic scope.

The amount of sites of this genre is expanding rapidly and adding the public release of the Google Maps application programming interface (API), with its use for creating *mashups* - web applications that gather information from various sources to create a new service of added value - this has resulted in a rapid increase of user-generated geographic data and resources.

Even Flickr, which is a photo-sharing site, provides surprising amounts of geospatial information due to the large amount of its members’ shared geotagged photos. For each geotagged photo they stored up to six different unique numeric identifiers that correspond to the hierarchy of places where a photo was taken: the neighborhood, the town, the county, and so on up to the continent. From that fact, they wondered if by taking all the photo information associated to a specific place ID would it be possible to have enough data in order to generate a mostly accurate contour of that place. From that reasoning they created shapefile data ranging from countries, states, cities and even neighborhoods. This was done using just the coordinates from the photos associated with that place’s ID[6].

This boom of user-generated geographic resources offers the possibility of having free, rich and variant sources to use for gazetteer creation, like in [42] where the authors experiment creating a gazetteer from the collection of Flickr’s geotagged photos; and also for using them as a means to maintain and expand a gazetteer in a semi-automated fashion [41].

Also, the thematic scope of VGI allows for creating gazetteers specialized to different themes (e.g. a gazetteer of hiking routes) and purposes - considering its nature as a product of a willing participant population, one can use them as a cheaper and faster means of obtaining reliable real-time changes in their environment (notice of new streets, changes in place names, progression of natural disasters), basically making use of a network of ”human sensors” [21]; also as a way to bring more detail to localized areas

[3] `http://www.esri.com`

[4] `http://maps.google.com`

[5] `http://www.openstreetmap.org`

[6] `http://code.flickr.com/blog/tag/shapefiles/`

(e.g. section of a city, neighborhood, park). Whatever the case, its usefulness is clear as a way to add more variety and quantity of data to gazetteers.

However, using VGI also raises concerns to consider regarding its quality and credibility. Since it is the product of people with barely no training or experience, it clearly poses issues of its spatial accuracy and because users of geographic information have developed expectations about the quality provided by traditional sources such as national mapping agencies and corporations, based on their experience, standards or reputation, the issue of VGI credibility is raised because of its provenance from people with no reputation or standards associated. Their information is merely asserted as being credible as opposed to the authoritative nature of those agencies and corporations.

Research regarding these concerns [22, 20] seems to point in the direction of studying the motivations and behavior of VGI providers as a way to establish credibility and even some research has been made to try and establish policies for gazetteer selection and approval of VGI sources regarding their trust in a source, based on its contributions, opinion from other users, previously approved information used from that source, etc [41].

3.4 Duplicate Detection

As mentioned in chapter 2, gazetteers are geospatial directories for named places. 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 [30], identity uncertainty [24], record linkage [63], among others [46, 56, 45]. 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* [43] (derived from the minimum number of character deletions, insertions or substitutions required to equate two strings)
- *Monge-Elkan distance* [46] (similar to the Levenshtein distance, but assigning a relatively lower cost to a sequence of insertions or deletions)
- *Jaro-Winkler metric* [63] (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)

Besides string similarity metrics, there has also been research addressing the computation of similarity scores between other types of data such as categorical information (information based on having a set of objects labeled with terms from vocabularies, thesauri, etc).

Similarity measures for computing scores for categorical information based on multi-sets of objects include:

- the *Jaccard coefficient* [36] measures similarity as the size of the intersection divided by the size of the union of the sample sets
- *Dice's coefficient* [16] also measures similarity between multi-sets of objects, and is defined as two times the size of the intersection divided by the sum of the sizes of the sample sets.

Both the Jaccard or the Dice coefficient can also be applied as string similarity metrics, by seeing the strings as sets of characters or even as sets of word tokens [14].

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 [54] and Support Vector Machines (SVMs) [39].

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 dissertation, 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 [57]. 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 [57]. 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 [55]. 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 [18]. 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 [30], canopy clustering [45] or filtering [64].

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.

Chapter 4

Problem Definition and Analysis

This chapter presents the motivation for the work developed for this dissertation and the requirements for the work. In it, an introduction in greater detail of the Europeana project, where the gazetteer will be used, is given and also its purpose in the project. Then, a definition of the work and requirements to be fulfilled are presented.

4.1 Motivation: The Initiative Europeana

The Europeana project was initiated from the European Commission's proposal for the creation of a virtual European library, aiming to make Europe's cultural and scientific resources accessible for all. The idea was creating a single information space or, in this case, a single search engine for browsing a collection of European digital libraries.

Digital libraries are organized collections of digital content made available to the public by cultural and scientific institutions (libraries, archives and museums) and private content holders (e.g. publishers). They can consist of all kinds of "physical" material that has been digitized (books, audiovisual material, photographs, documents in archives etc) and material originally produced in digital format.

Originally known as the European digital library network - EDLnet - Europeana is a partnership of over 100 representatives of heritage and knowledge organizations and IT experts from throughout Europe. These experts contribute to enhance and provide core products for service operation. There are also contributors - organizations that are providing content, some of which are aggregators (collect material from a range of other contributors, display it on their own website and also channel it into Europeana). These include the *Rijksmuseum* in Amsterdam, the *British Library* in London and the *Louvre* in Paris. While it is still in a development stage, a prototype is already available for use and version 1.0 should be released still in 2010 with access to over 10 million digital objects.

As mentioned, there are several IT experts and organizations that help develop and enhance the Europeana service. These experts and organizations are organized into a network called Europeana Connect. The purpose of this network aims to deliver core components which are essential for the development and enhancement of Europeana, so as to become a truly interoperable, multilingual and user-oriented service. More specifically, they intend to, among other aspects:

- **Provide multilingual searching and browsing** including translation tools and other language resources to enable multilingual searching of objects and data in Europeana.

- **Semantically enrich digital content in Europeana** creating new connections between objects. Semantic enrichment will make Europeana content more accessible, reusable and exploitable.

- **Enhance usability and functionality** by integrating multimedia annotation (digital books, images and other content); also by integrating geographical information services so as to provide more power to the search tools by adding geographical information about the digital objects.

- **Deploy key infrastructure components for Europeana** including an OAI (The OAI - Open Archives Initiative[1] - develops and promotes interoperability standards aimed at facilitating the efficient dissemination of content) Management Infrastructure to handle large-scale metadata harvesting, a Metadata Registry to ensure interoperability, a Service Registry to enable integration of external added-value services and a Resolution Discovery Service to allow unique resource identification.

The network involves a work force comprised of people from several different countries, each organized into different work packages associated with particular achievements or deliverables. Also, besides the inherent complexities of a project of this magnitude with a distributed work force, there is the added challenge of synchronizing the project with the requirements and planning of the Europeana service. This means delivering both according to the project's Description of Work and Europeana's release planning and requirements. And, unlike other projects, the deliverables should not be prototypes but production-ready components, so as to integrate them directly into the Europeana service. So, the measure of success from this project will be measured by the success in delivering sustainable quality products for integrating into Europeana.

In this context, the work performed for the purposes of this dissertation fits into the third of the points mentioned above. The objective is to further enrich the content provided by Europeana by adding information about geographic locations featured in the digital objects. This is accomplished by the use of two 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's job is only that - detecting geographic references and adding them as new elements of information into the records. It 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.

It serves as a complementary service with the Geoparser, receiving requests of geographic entities from the Geoparser and responding with all the pertinent information regarding those geographic entities. It does this through the use of the ADL Gazetteer Protocol described in Chapter 2.

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.

4.2 Requirements

Having established the client for the work of this dissertation, it is now necessary to gather the requirements for the gazetteer service. This section presents both the non-functional and functional requirements

[1] http://www.openarchives.org/

for the Gazetteer system. The functional requirements are described using "use case diagrams" as defined by the Unified Modeling Language (UML)[2], which is a well-known language that allows for simplified understanding by most developers familiar with modern software design. A use case is a set of scenarios describing an interaction between one or more actors and a system. Use cases can be described in diagrams, and detailed using textual descriptions. An actor in a use case represents any entity interacting with the system, which can be a human user or another system.

4.2.1 Non-Functional Requirements

Non-functional requirements define important constraints of the system. They describe the "how", "when" and "where" of the system, that do not depend on the use cases but on the businesses and technical constraints.

R-1. Operational Requirements	
1.1	The system must be as portable as possible, minimizing any kind of platform-specific dependencies.
1.2	The system must provide simple service interfaces in web service form, preferably according to the Representational State Transfer (REST) principles.
1.3	The system must provide all the necessary and useful operations for management of both the gazetteer system and its content.

Table 4.1: Operational Requirements

R-2. Interface Requirements	
2.1	The administrative interface must have a short learning curve and non unnecessary technical terms, to be allow use by non technical staff using the Gazetteer.
2.2	The user interface must provide all the necessary information to support the user interaction with the system.
2.3	The user interface for creating features must provide a practical and preferably visual way for users to identify locations in the world map so that its creation is quicker and less error-prone.
2.4	The administrative interface must only offer access to both information and functionality for users with the necessary credentials.

Table 4.2: Interface Requirements

[2] http://www.uml.org

Chapter 5

Proposed Solution

This chapter describes the planned design and architecture for the Gazetteer System. It shows the component architecture of the system and describes the responsibility of each component, along with connections to external services and interfaces.

5.1 System Architecture

The Gazetteer System Architecture is composed of three internal subsystems (see Figure 5.1): the *Gazetteer User Interface*, the *Gazetteer Service*, and a *Repository.*

The **Gazetteer User Interface** is responsible for supporting all processes that require user interaction with Gazetteer, mainly for discovery and access of Features and management. It was separated from the Gazetteer Service since searching and browsing for Geographic Features may require customized user interfaces (e.g. geographic browsing, hierarchical browsing, searching given a bounding box). This way the Gazetteer User Interface may be designed, tested and improved (whether in functionality or with better user interaction) separately from the Gazetteer Service, which needs to be stable. This subsystem is comprised of two UI : a public one designed for browsing and searching Features and a restricted one designed for administrating the system and the features in the Repository.

The **Gazetteer Service** is responsible for supporting all the Gazetteer processes that require accessing the Feature data stored in the Repository. It serves as a simple service providing the basic functionalities of a ISO and OGC defined gazetteer service, instead interacting with an “ecology” of services that provide additional functionalities to the service (such as importing new content, validating existing content, exporting content, administration functions, etc.) The next section will detail the contents of this subsystem in terms of components and interacting services.

The **Repository** is a set of databases responsible for storing and structuring the Gazetteer Feature data. The structure for storing the data should consist of two databases: one with tables serving as indexes for each searchable criteria of a Feature stored in the Gazetteer; the other one is the storage database, containing the full records of the Gazetteer features.

There are also external services that interact with the Gazetteer, such as an Ingester, a Duplicate Detector and Exporter services, which will be discussed in the following section.

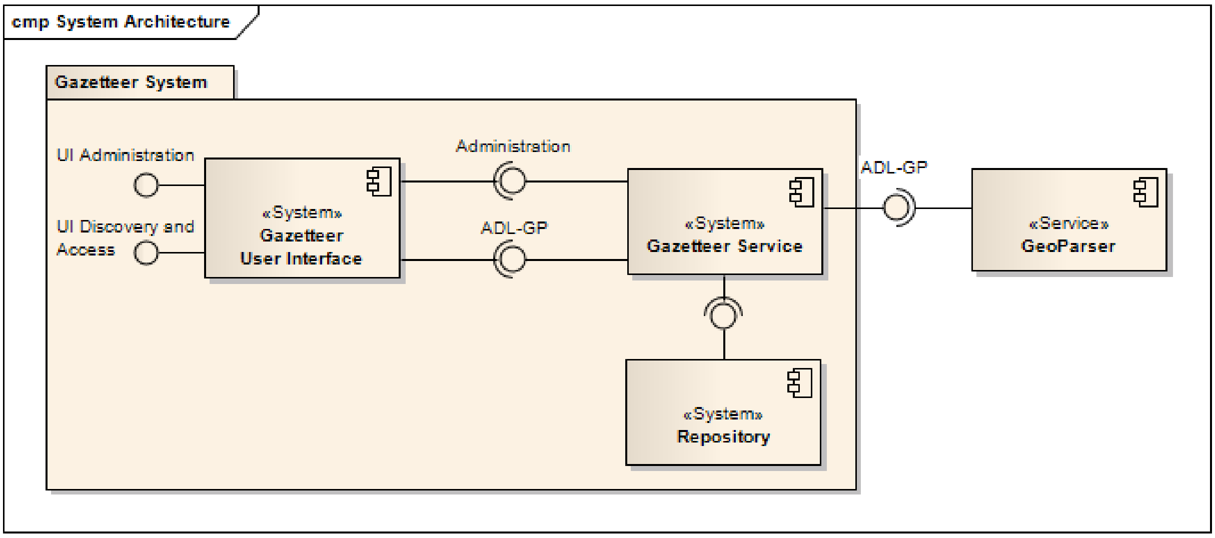


Figure 5.1: Overview of the system architecture for the Gazetteer System.

5.2 Component Architecture

This architecture shows the component parts that make up the system. As described in the previous section the system is decomposed in three other subsystems, the Gazetteer User Interface, the Gazetteer Service and the Repository. The Gazetteer User Interface is composed of two components: *the Public UI, Administration UI;* while the Gazetteer Service is comprised of two internal components: *Search Support and Storage Manager*. This Gazetteer Service borrows additional functionality from other external services such as the *Ingester*, the *Duplicate Detector* and the *Exporter*, which are all invoked by an *Administration Manager* component. Figure 5.2 presents an overview of the components of the subsystem, the external services and their relationships with each other, while the following points describe in detail each of these components.

Component: *Public User Interface*

Description: Responsible for providing all the operations which require user interaction for browsing and searching for Features in the Gazetteer.

Dependencies: *Search Support*: Requires this component's operations to answer a user for feature searches

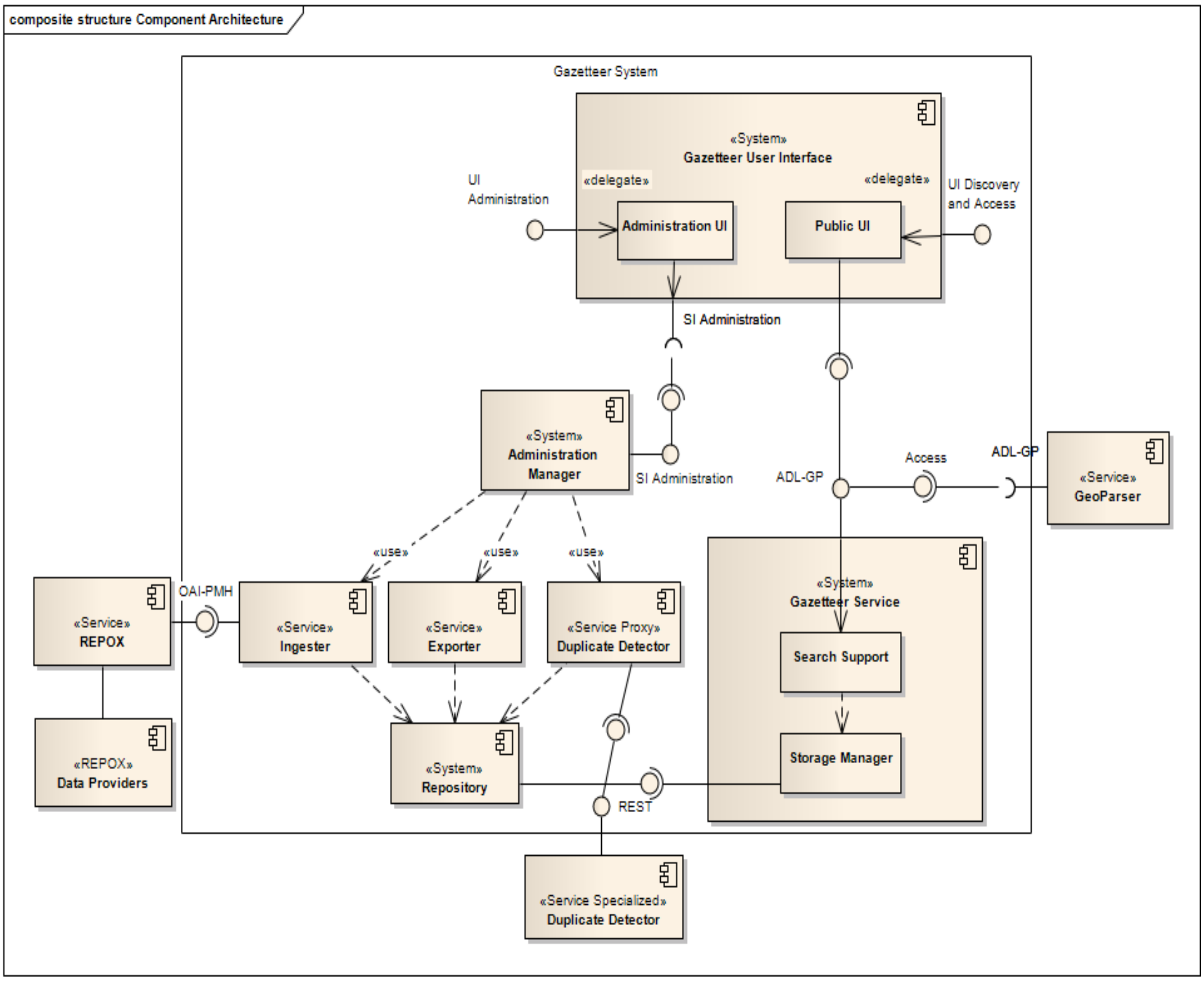


Figure 5.2: Overview of the component architecture for the Gazetteer System.

Component: *Administrative User Interface*

Description: Responsible for providing all the operations which require user interaction for managing the users, features and overall system health for the Gazetteer

Dependencies: *Administration Manager*: Requires this component's operations to allow access to the system.

Component: *Administration Manager*

Description: Responsible for providing all operations that are related with the management of users, maintenance of the system and also provides operations related with the health of stored features, such as report generation and editing features. Is restricted to users with administrator or registered user privileges.

Dependencies: *Storage Manager*: For access to the repository.

Service: *Ingester*

Description: Responsible for providing operations related to the adding an updating of features in the Gazetteer. The planned implementation has the Ingester as registered user of the REPOX service which will be responsible for aggregating data providers for the Gazetteer, leaving to the Ingester the sole task of querying the service for new content, fetching it and adding it to the Gazetteer.

Dependencies: *Repository*: For access to the gazetteer content.

Component: *Search Support*

Description: Responsible for supporting all operations that require querying the repository for Features.

Dependencies: *Storage Manager*: For access to the repository.

Service: *Exporter*

Description: Responsible for providing operations for exporting or publishing the data contained in the repository.

Dependencies: *Repository*: For access to the gazetteer content.

Service: "System Proxy"*Duplicate Checker*

Description: Responsible for scanning a collection of Features for duplicates. Can be used both to compare new entries from the Ingester to the existing collection or to check the Repository in case of changes to existing Features. This proxy provides an abstract operation "Duplicate Detection" that can be resolved by using an array of specialized services which provide different degrees of depth and complexity in terms of the duplicate detection, ranging from basic algorithms which can be scheduled regularly to more complete and extensive methods which are time consuming and can be performed only once a month or similar.

Dependencies: *Repository*: For access to the gazetteer content.

Component: *Storage Manager*

Description: Responsible for managing the system access to the data repository where the features are stored.

Dependencies: *Repository*: For access managing.

Chapter 6

Implementation

This chapter serves for explaining how the system is actually implemented and how it differs from the proposed architecture.

6.1 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:

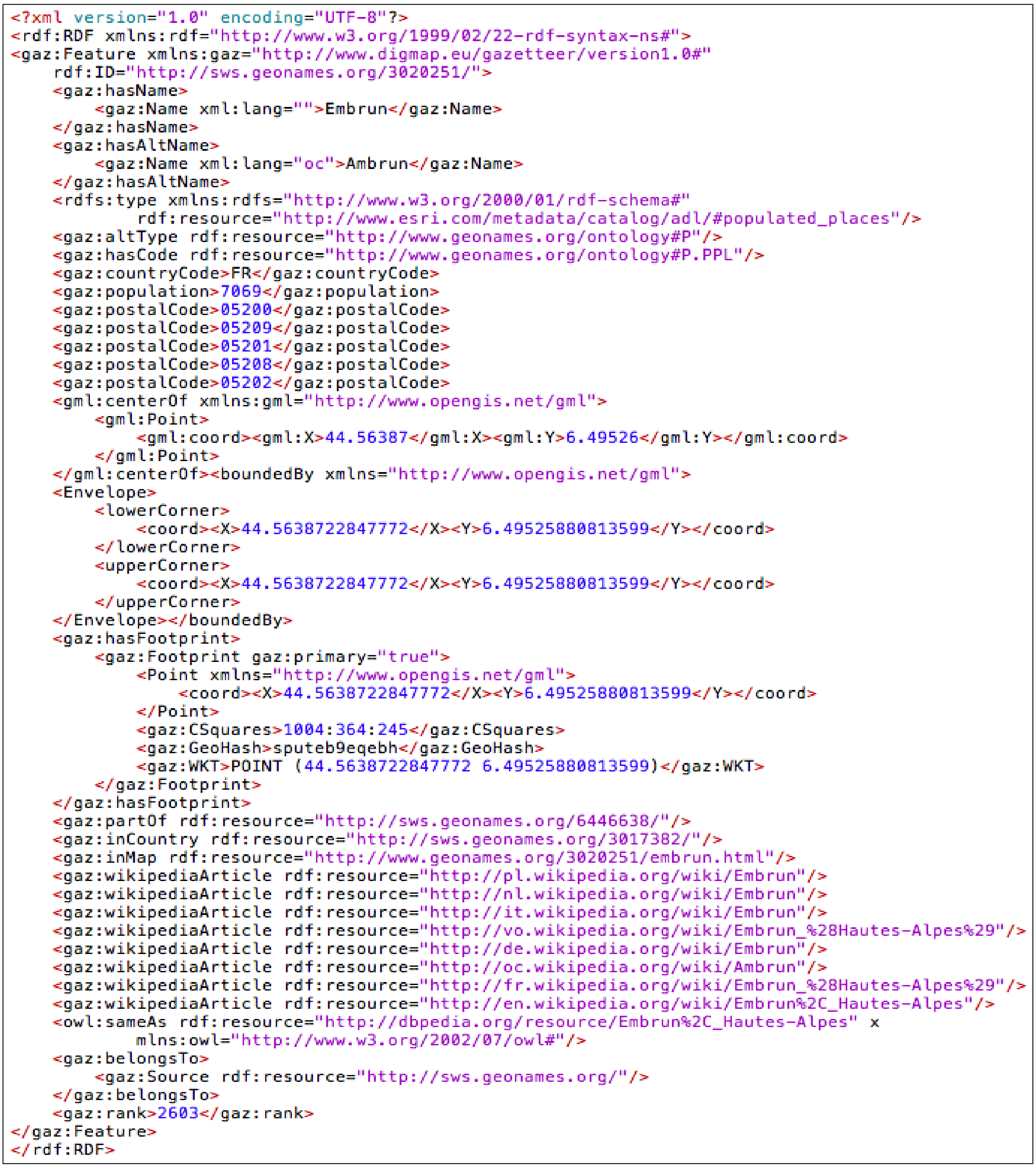
- ID - both an internal and source ID are stored for each feature;
- primary name associated with the feature;
- one or more alternate names, which may have a language code associated with them;
- spatial footprints or temporal coverages, depending on the feature;
- one or more feature types;
- relationships 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



```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
<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_places"/>
    <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:coord>
        </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/"/>
    <gaz:inMap rdf:resource="http://www.geonames.org/3020251/embrun.html"/>
    <gaz:wikipediaArticle rdf:resource="http://pl.wikipedia.org/wiki/Embrun"/>
    <gaz:wikipediaArticle rdf:resource="http://nl.wikipedia.org/wiki/Embrun"/>
    <gaz:wikipediaArticle rdf:resource="http://it.wikipedia.org/wiki/Embrun"/>
    <gaz:wikipediaArticle rdf:resource="http://vo.wikipedia.org/wiki/Embrun_%28Hautes-Alpes%29"/>
    <gaz:wikipediaArticle rdf:resource="http://de.wikipedia.org/wiki/Embrun"/>
    <gaz:wikipediaArticle rdf:resource="http://oc.wikipedia.org/wiki/Ambrun"/>
    <gaz:wikipediaArticle rdf:resource="http://fr.wikipedia.org/wiki/Embrun_%28Hautes-Alpes%29"/>
    <gaz:wikipediaArticle rdf:resource="http://en.wikipedia.org/wiki/Embrun%2C_Hautes-Alpes"/>
    <owl:sameAs rdf:resource="http://dbpedia.org/resource/Embrun%2C_Hautes-Alpes" x
            mlns:owl="http://www.w3.org/2002/07/owl#"/>
    <gaz:belongsTo>
        <gaz:Source rdf:resource="http://sws.geonames.org/"/>
    </gaz:belongsTo>
    <gaz:rank>2603</gaz:rank>
</gaz:Feature>
</rdf:RDF>
```

Figure 6.1: Internal Representation for a Gazetteer Feature of Embrun

output formats, through the use of Extensible Stylesheet Language Transformations (XSLT) sheets for transforming the data.

An example of a feature's internal representation is shown in Figure 6.1.

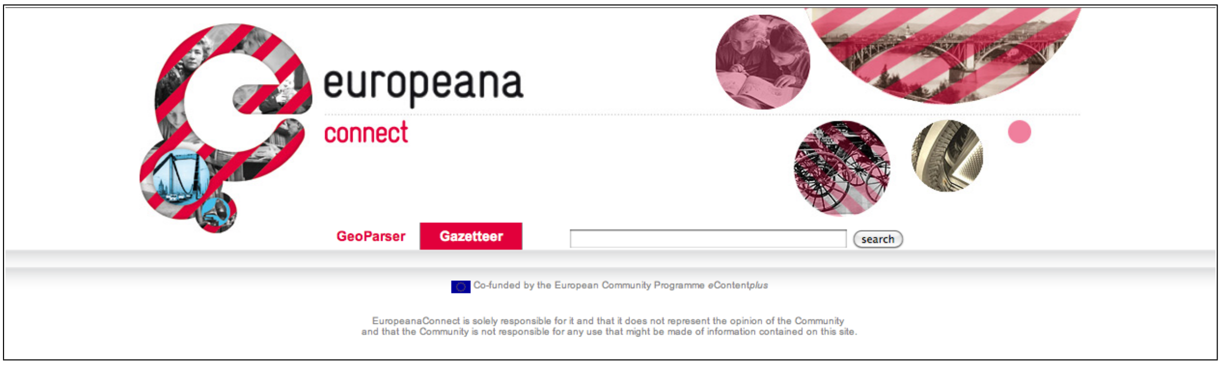
In this figure, one can see the feature description beginning with the **Feature** node, with an attribute that provides the internal identifier for this particular feature. The primary name is provided in the **hasName** node, while any alternative names are placed inside the **hasAltName** node. In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

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, followed by some links to external information sources that can aid in better describing the feature.

6.2 Gazetteer Human Public Interface



Co-funded by the European Community Programme eContent*plus*

EuropeanaConnect is solely responsible for it and that it does not represent the opinion of the Community and that the Community is not responsible for any use that might be made of information contained on this site.

Figure 6.2: Europeana Gazetteer Main page

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.

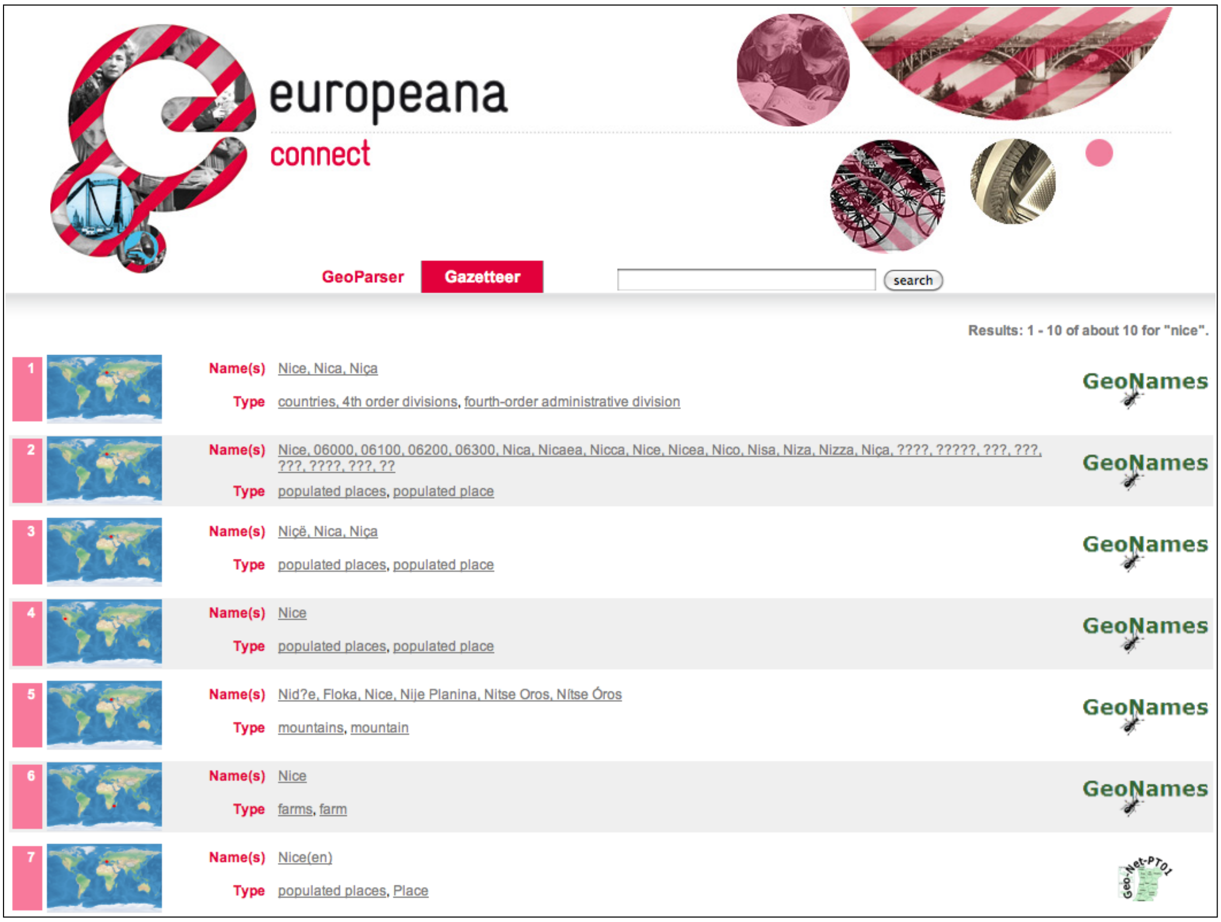
From its main page, shown in Figure 6.2, a user can perform a search for a feature name, for which any matching features are display in a result list - shown in Figure 6.3.

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.

Figure 6.4 shows the descriptive information portion of the description page. 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.



Results: 1 - 10 of about 10 for "nice".

1 Name(s) Nice, Nica, Niça
Type countries, 4th order divisions, fourth-order administrative division

2 Name(s) Nice, 06000, 06100, 06200, 06300, Nica, Nicaea, Nicca, Nice, Nicea, Nico, Nisa, Niza, Nizza, Niça, ????, ?????, ???, ???, ???, ????, ???, ??
Type populated places, populated place

3 Name(s) Niçë, Nica, Niça
Type populated places, populated place

4 Name(s) Nice
Type populated places, populated place

5 Name(s) Nid?e, Floka, Nice, Nije Planina, Nitse Oros, Nítse Óros
Type mountains, mountain

6 Name(s) Nice
Type farms, farm

7 Name(s) Nice(en)
Type populated places, Place

Figure 6.3: Gazetteer search list excerpt for the term “**Nice**”

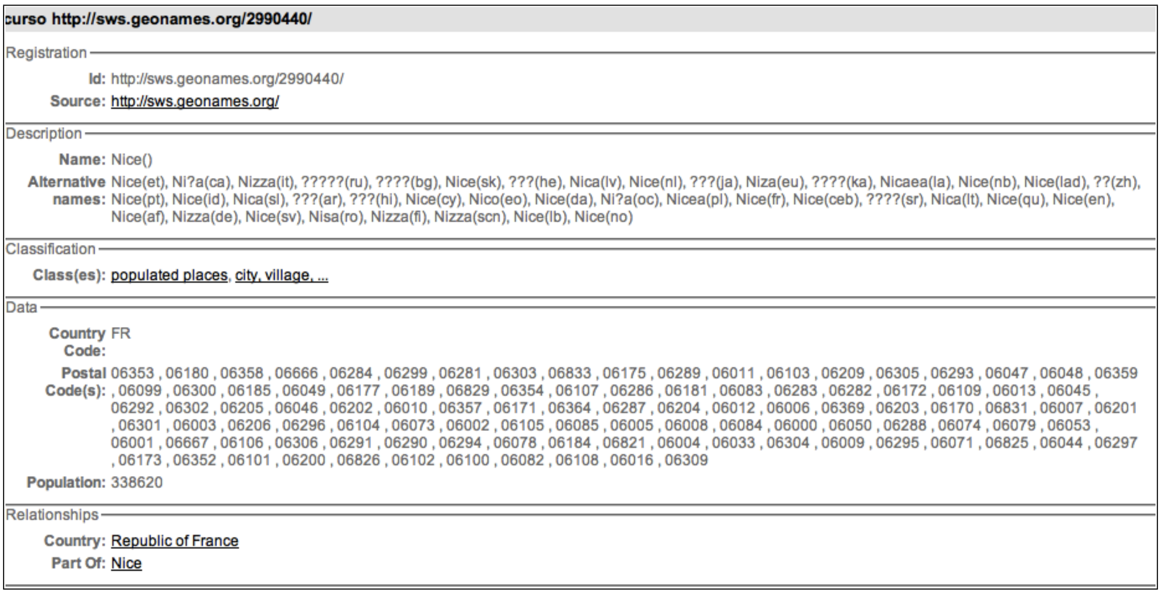
Figure 6.5 shows the geographic information portion, where one can see both a visual representation of the feature and coordinate representations;

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.

6.3 Gazetteer Service

The core Gazetteer Service is constructed as an independent build of Java[1] projects. It is built upon the initial one used for the purposes of the DIGMAP project. The main remnants of that initial build are the Storage Manager and Search Support components. They provide the basis for enabling access and retrieval of features stored in the repository. The Search Support builds mostly on the use of the ADL Gazetteer Protocol, using it as the means to structure queries to process and for the Storage Manager to return a list of feature results. The gazetteer is setup in this way in order to be used as a simple service, providing the basic functionalities established by ISO and OGC standards. Any other functionality is “borrowed” by way of the Administration Manager Service, that oversees an “ecology” of external services, each of them serving a specific purpose (the Ingester for importing new data from other providers, the Duplicate Detection service for controlling the overall quality of the existing gazetteer data and the Exporter for providing means to extract data from the Gazetteer for publishing or other purposes).

[1] `http://www.java.com`



curso http://sws.geonames.org/2990440/

Registration

Id: http://sws.geonames.org/2990440/
Source: http://sws.geonames.org/

Description

Name: Nice()
Alternative names: Nice(et), Ni?a(ca), Nizza(it), ?????(ru), ????(bg), Nice(sk), ???(he), Nica(lv), Nice(nl), ???(ja), Niza(eu), ????(ka), Nicaea(la), Nice(nb), Nice(lad), ??(zh), Nice(pt), Nice(id), Nica(sl), ???(ar), ???(hi), Nice(cy), Nico(eo), Nice(da), Ni?a(oc), Nicea(pl), Nice(fr), Nice(ceb), ????(sr), Nica(lt), Nice(qu), Nice(en), Nice(af), Nizza(de), Nice(sv), Nisa(ro), Nizza(fi), Nizza(scn), Nice(lb), Nice(no)

Classification

Class(es): populated places, city, village, ...

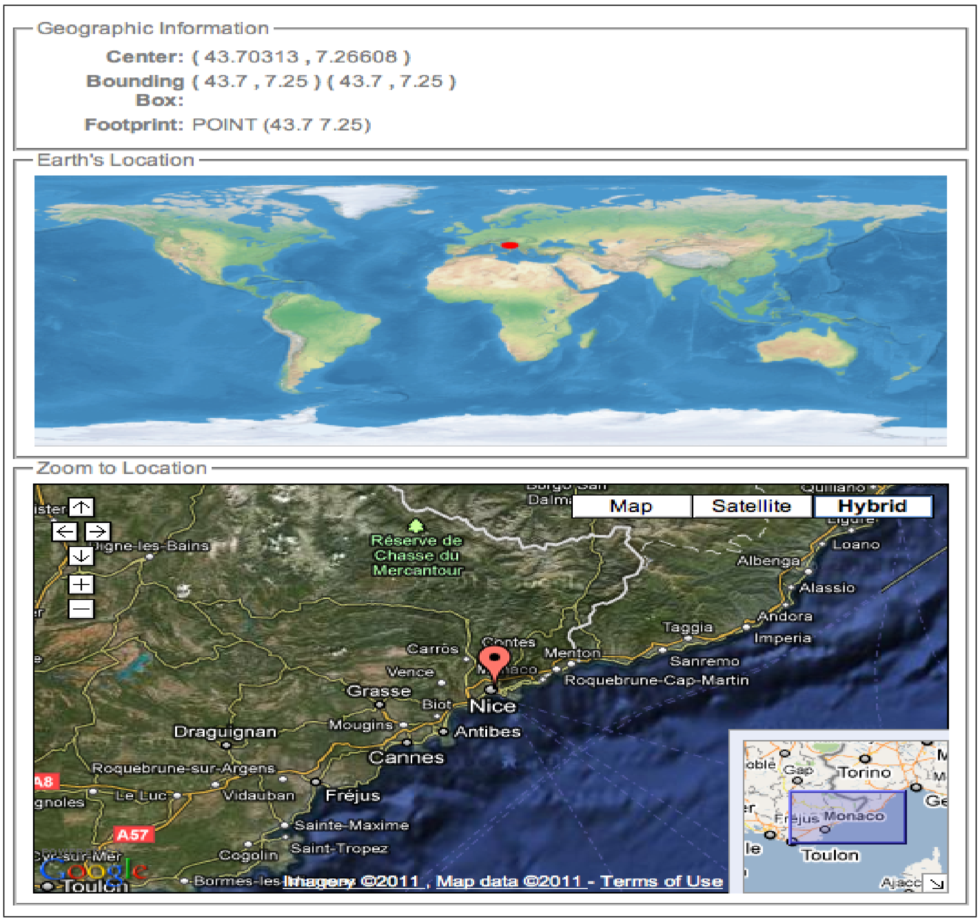
Data

Country Code: FR
Postal Code(s): 06353 , 06180 , 06358 , 06666 , 06284 , 06299 , 06281 , 06303 , 06833 , 06175 , 06289 , 06011 , 06103 , 06209 , 06305 , 06293 , 06047 , 06048 , 06359 , 06099 , 06300 , 06185 , 06049 , 06177 , 06189 , 06829 , 06354 , 06107 , 06286 , 06181 , 06083 , 06283 , 06282 , 06172 , 06109 , 06013 , 06045 , 06292 , 06302 , 06205 , 06046 , 06202 , 06010 , 06357 , 06171 , 06364 , 06287 , 06204 , 06012 , 06006 , 06369 , 06203 , 06170 , 06831 , 06007 , 06201 , 06301 , 06003 , 06206 , 06296 , 06104 , 06073 , 06002 , 06105 , 06085 , 06005 , 06008 , 06084 , 06000 , 06050 , 06288 , 06074 , 06079 , 06053 , 06001 , 06667 , 06106 , 06306 , 06291 , 06290 , 06294 , 06078 , 06184 , 06821 , 06004 , 06033 , 06304 , 06009 , 06295 , 06071 , 06825 , 06044 , 06297 , 06173 , 06352 , 06101 , 06200 , 06826 , 06102 , 06100 , 06082 , 06108 , 06016 , 06309
Population: 338620

Relationships

Country: Republic of France
Part Of: Nice

Figure 6.4: Description page for feature of “**Nice**” - Focus on textual information



Geographic Information

Center: (43.70313 , 7.26608)
Bounding Box: (43.7 , 7.25) (43.7 , 7.25)
Footprint: POINT (43.7 7.25)

Earth's Location

Zoom to Location

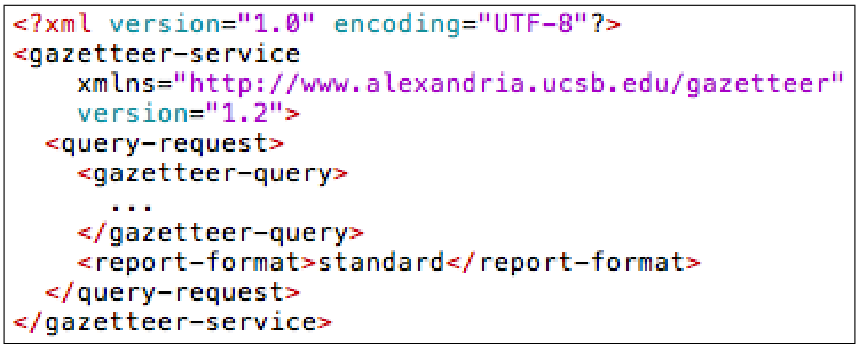
Figure 6.5: Description page for feature of “**Nice**” - Focus on geographic information

6.3.1 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 6.6. The query type(s) can then be specified inside the **query-request** node.

Query Types

The following section describes the types of queries the system supports and some 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 6.6: 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" />`

D: *Relationship Query*

Description: This query type fetches the features that share a specified relation type with a specified feature internal identifier. The relations covered are:

“*partOf*”: The feature is contained within the area of target feature;

“*contains*”: The feature contains target feature;

“*adjacentTo*”: The feature has a frontier or border with target feature;

“*inCountry*”: The feature is part of target country feature;

“*inContinent*”: The feature is part of target continent feature;

Example:
```
<relationship-query relation="inCountry"
  target-identifier="http://sws.geonames.org/2990440/"/>
```

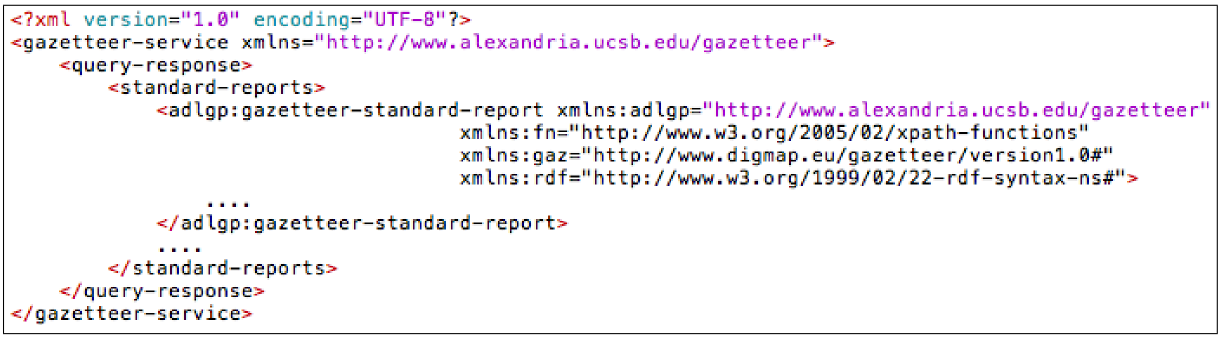
E: *Combine Queries*

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>
```

Query Response

The XML response for any query request type consists of a list of all search matches to the query parameters. Figure 6.7 depicts the base format of a query response:



```
<?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>
```

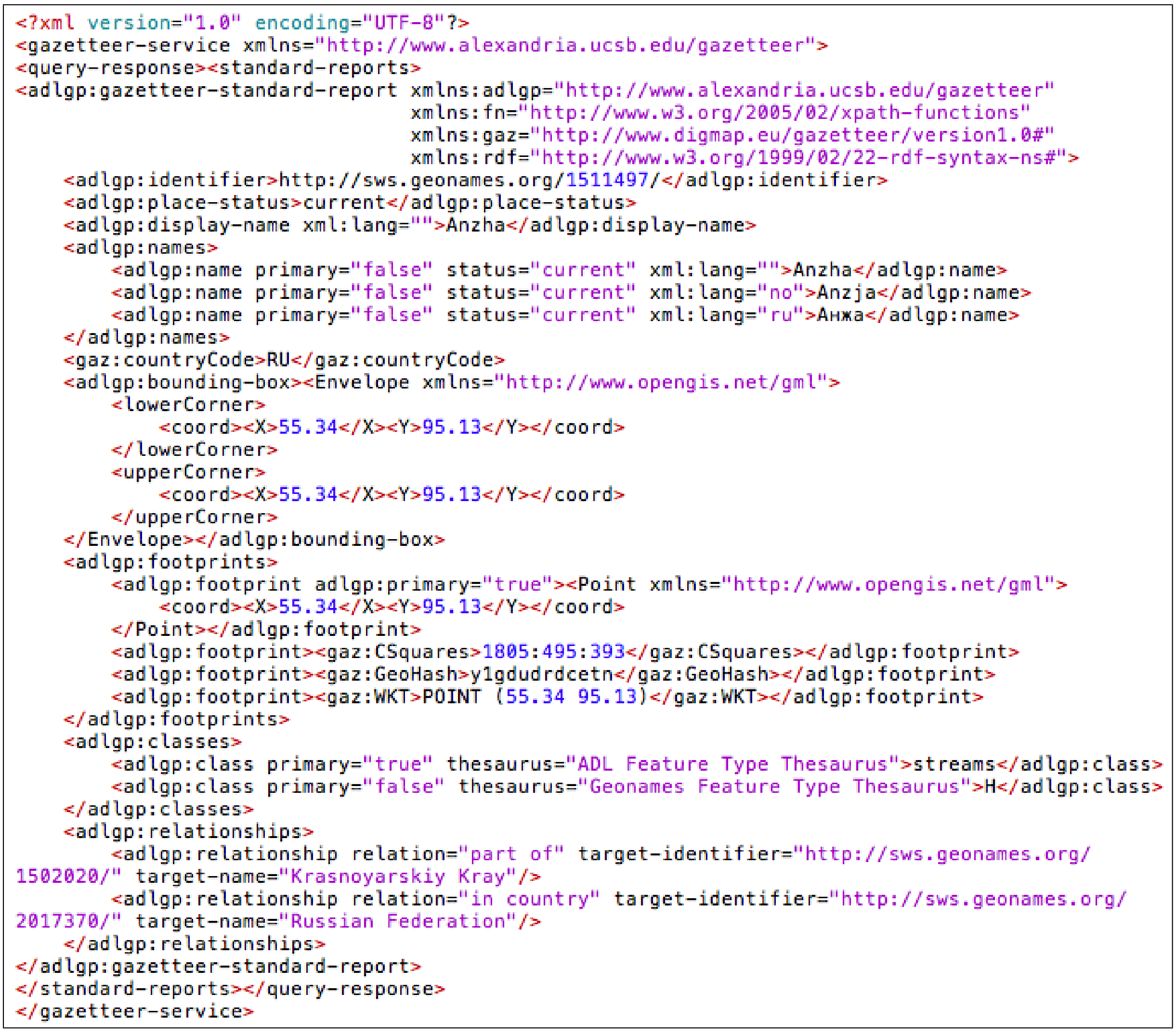
Figure 6.7: Base format for a gazetteer feature query response

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[2], 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 6.8:



```
<?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:identifier>http://sws.geonames.org/1511497/</adlgp:identifier>
    <adlgp:place-status>current</adlgp:place-status>
    <adlgp:display-name xml:lang="">Anzha</adlgp:display-name>
    <adlgp:names>
        <adlgp:name primary="false" status="current" xml:lang="">Anzha</adlgp:name>
        <adlgp:name primary="false" status="current" xml:lang="no">Anzja</adlgp:name>
        <adlgp:name primary="false" status="current" xml:lang="ru">Анжа</adlgp:name>
    </adlgp:names>
    <gaz:countryCode>RU</gaz:countryCode>
    <adlgp:bounding-box><Envelope xmlns="http://www.opengis.net/gml">
        <lowerCorner>
            <coord><X>55.34</X><Y>95.13</Y></coord>
        </lowerCorner>
        <upperCorner>
            <coord><X>55.34</X><Y>95.13</Y></coord>
        </upperCorner>
    </Envelope></adlgp:bounding-box>
    <adlgp:footprints>
        <adlgp:footprint adlgp:primary="true"><Point xmlns="http://www.opengis.net/gml">
            <coord><X>55.34</X><Y>95.13</Y></coord>
        </Point></adlgp:footprint>
        <adlgp:footprint><gaz:CSquares>1805:495:393</gaz:CSquares></adlgp:footprint>
        <adlgp:footprint><gaz:GeoHash>y1gdudrdcetn</gaz:GeoHash></adlgp:footprint>
        <adlgp:footprint><gaz:WKT>POINT (55.34 95.13)</gaz:WKT></adlgp:footprint>
    </adlgp:footprints>
    <adlgp:classes>
        <adlgp:class primary="true" thesaurus="ADL Feature Type Thesaurus">streams</adlgp:class>
        <adlgp:class primary="false" thesaurus="Geonames Feature Type Thesaurus">H</adlgp:class>
    </adlgp:classes>
    <adlgp:relationships>
        <adlgp:relationship relation="part of" target-identifier="http://sws.geonames.org/
1502020/" target-name="Krasnoyarskiy Kray"/>
        <adlgp:relationship relation="in country" target-identifier="http://sws.geonames.org/
2017370/" target-name="Russian Federation"/>
    </adlgp:relationships>
</adlgp:gazetteer-standard-report>
</standard-reports></query-response>
</gazetteer-service>
```

Figure 6.8: Example query response

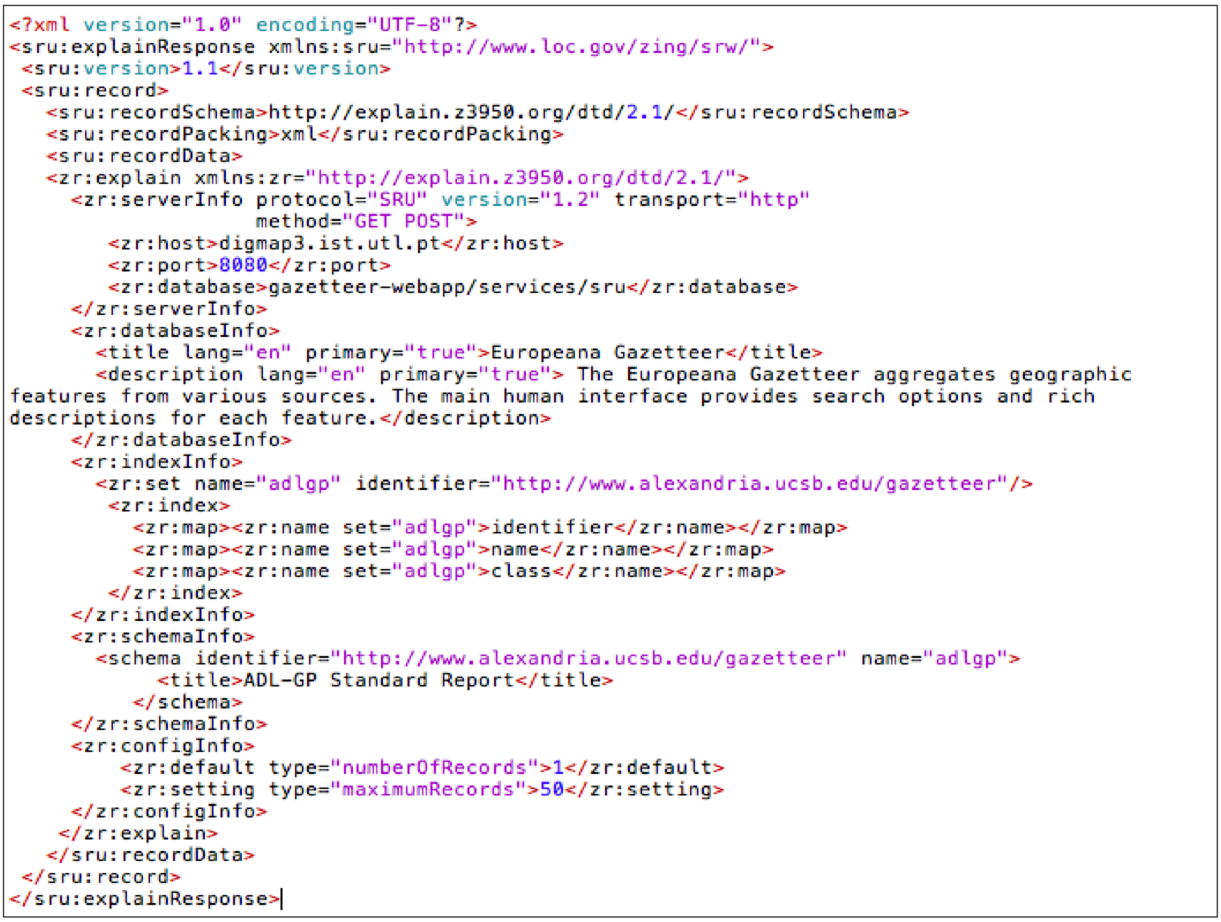
[2] http://www.iso.org/iso/country_codes.htm

6.3.2 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) - 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. A detailed explanation of the service is provided in chapter 2.

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.

The explain file for the current version of the interface is shown in Figure 6.9.



```
<?xml version="1.0" encoding="UTF-8"?>
<sru:explainResponse xmlns:sru="http://www.loc.gov/zing/srw/">
 <sru:version>1.1</sru:version>
 <sru:record>
   <sru:recordSchema>http://explain.z3950.org/dtd/2.1/</sru:recordSchema>
   <sru:recordPacking>xml</sru:recordPacking>
   <sru:recordData>
   <zr:explain xmlns:zr="http://explain.z3950.org/dtd/2.1/">
     <zr:serverInfo protocol="SRU" version="1.2" transport="http"
                    method="GET POST">
        <zr:host>digmap3.ist.utl.pt</zr:host>
        <zr:port>8080</zr:port>
        <zr:database>gazetteer-webapp/services/sru</zr:database>
     </zr:serverInfo>
     <zr:databaseInfo>
       <title lang="en" primary="true">Europeana Gazetteer</title>
       <description lang="en" primary="true"> The Europeana Gazetteer aggregates geographic
features from various sources. The main human interface provides search options and rich
descriptions for each feature.</description>
     </zr:databaseInfo>
     <zr:indexInfo>
       <zr:set name="adlgp" identifier="http://www.alexandria.ucsb.edu/gazetteer"/>
        <zr:index>
          <zr:map><zr:name set="adlgp">identifier</zr:name></zr:map>
          <zr:map><zr:name set="adlgp">name</zr:name></zr:map>
          <zr:map><zr:name set="adlgp">class</zr:name></zr:map>
        </zr:index>
     </zr:indexInfo>
     <zr:schemaInfo>
       <schema identifier="http://www.alexandria.ucsb.edu/gazetteer" name="adlgp">
            <title>ADL-GP Standard Report</title>
          </schema>
     </zr:schemaInfo>
     <zr:configInfo>
         <zr:default type="numberOfRecords">1</zr:default>
         <zr:setting type="maximumRecords">50</zr:setting>
     </zr:configInfo>
    </zr:explain>
   </sru:recordData>
 </sru:record>
</sru:explainResponse>
```

Figure 6.9: Explain record

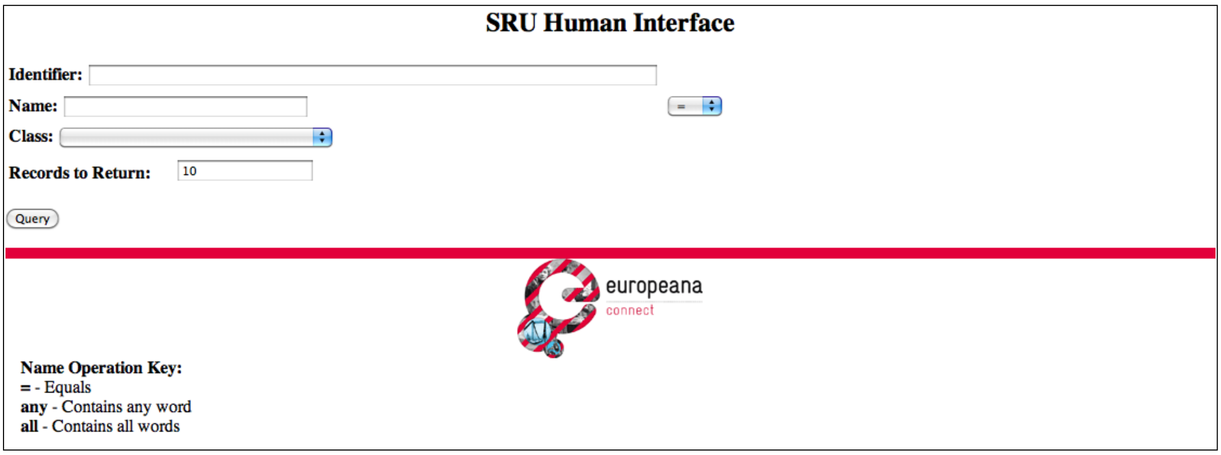
From the explain record, one can obtain information regarding the server and database where the protocol is running, the supported indexes for search (identifier, name and class) , the response schema (ADL-GP Standard Report) and some configuration options.

Human Test Interface

For testing purposes and also as a way to assist users unfamiliar with the protocol, a Human Interface is provided. It can be accessed through the URL: `http://digmap3.ist.utl.pt:8080/gazetteer-webapp/sru.jsp`

At present, the interface supports queries by internal identifier or by name and/or class of a feature.

Both the identifier and name fields are text boxes where a user can type the desired id or name and the class field is a drop list of all the attributed feature types present in the gazetteer. Also, the number of records to return can be defined. Default is set to 10 records. Figure 6.10 demonstrates this.



SRU Human Interface

Identifier:

Name: =

Class:

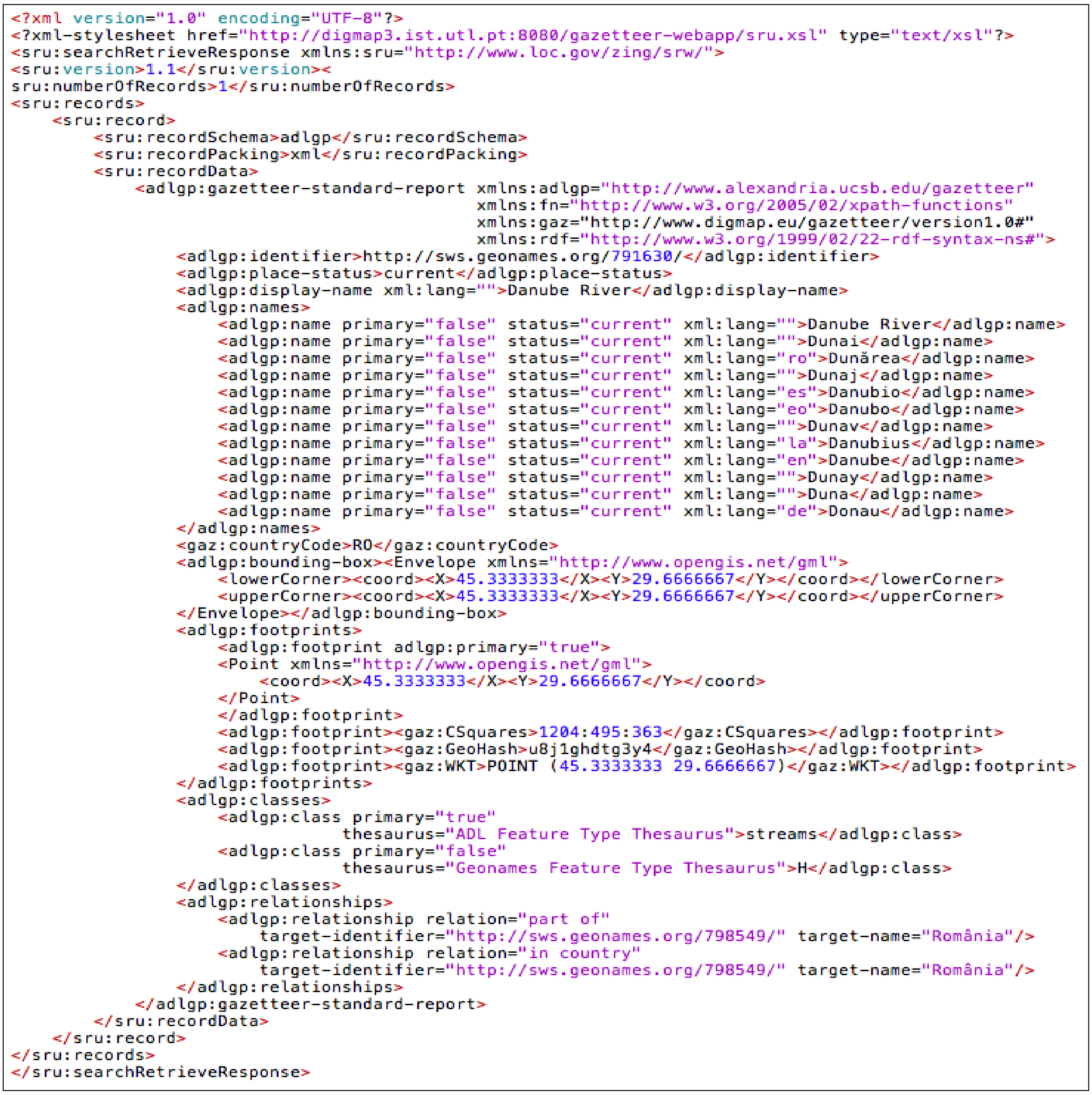
Records to Return: 10

Name Operation Key:
= - Equals
any - Contains any word
all - Contains all words

Figure 6.10: SRU Human Test Interface

Once the query is submitted, a list of records is returned matching the query parameters. For example for a query for features named Danube and class of streams returns the XML result list in Figure 6.11.

That list corresponds to the following HTML representation, accomplished by applying a stylesheet to the XML result file. Only some fields are selected are presented in this fashion (Figure 6.12). To note that the field *numberOfRecords*, represented in the Total Records Found line, presents the full number of records that matched the query. This value is always equal or less than the number of records defined in the query interface.



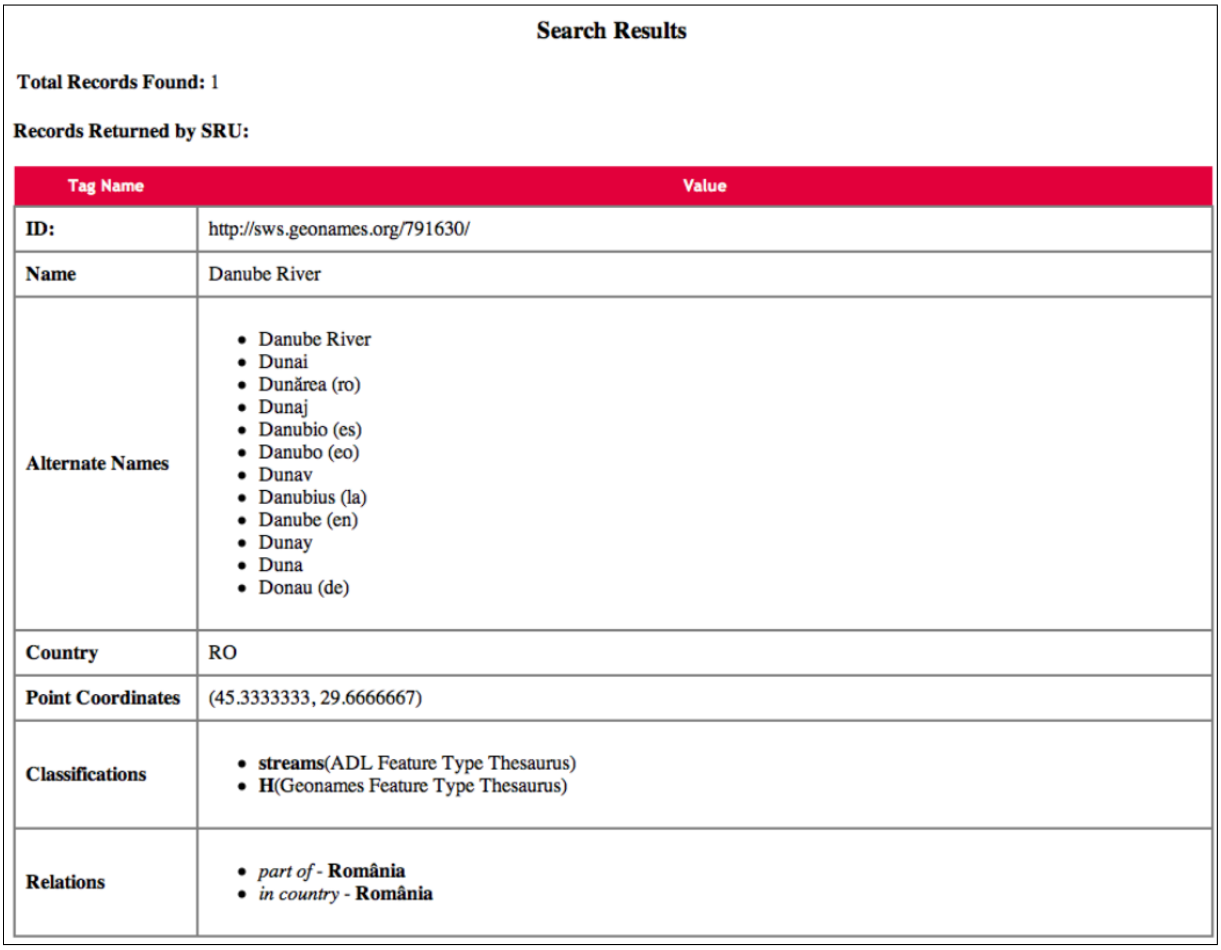
```xml
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet href="http://digmap3.ist.utl.pt:8080/gazetteer-webapp/sru.xsl" type="text/xsl"?>
<sru:searchRetrieveResponse xmlns:sru="http://www.loc.gov/zing/srw/">
<sru:version>1.1</sru:version><
sru:numberOfRecords>1</sru:numberOfRecords>
<sru:records>
    <sru:record>
        <sru:recordSchema>adlgp</sru:recordSchema>
        <sru:recordPacking>xml</sru:recordPacking>
        <sru:recordData>
            <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:identifier>http://sws.geonames.org/791630/</adlgp:identifier>
                <adlgp:place-status>current</adlgp:place-status>
                <adlgp:display-name xml:lang="">Danube River</adlgp:display-name>
                <adlgp:names>
                    <adlgp:name primary="false" status="current" xml:lang="">Danube River</adlgp:name>
                    <adlgp:name primary="false" status="current" xml:lang="">Dunai</adlgp:name>
                    <adlgp:name primary="false" status="current" xml:lang="ro">Dunărea</adlgp:name>
                    <adlgp:name primary="false" status="current" xml:lang="">Dunaj</adlgp:name>
                    <adlgp:name primary="false" status="current" xml:lang="es">Danubio</adlgp:name>
                    <adlgp:name primary="false" status="current" xml:lang="eo">Danubo</adlgp:name>
                    <adlgp:name primary="false" status="current" xml:lang="">Dunav</adlgp:name>
                    <adlgp:name primary="false" status="current" xml:lang="la">Danubius</adlgp:name>
                    <adlgp:name primary="false" status="current" xml:lang="en">Danube</adlgp:name>
                    <adlgp:name primary="false" status="current" xml:lang="">Dunay</adlgp:name>
                    <adlgp:name primary="false" status="current" xml:lang="">Duna</adlgp:name>
                    <adlgp:name primary="false" status="current" xml:lang="de">Donau</adlgp:name>
                </adlgp:names>
                <gaz:countryCode>RO</gaz:countryCode>
                <adlgp:bounding-box><Envelope xmlns="http://www.opengis.net/gml">
                    <lowerCorner><coord><X>45.3333333</X><Y>29.6666667</Y></coord></lowerCorner>
                    <upperCorner><coord><X>45.3333333</X><Y>29.6666667</Y></coord></upperCorner>
                </Envelope></adlgp:bounding-box>
                <adlgp:footprints>
                    <adlgp:footprint adlgp:primary="true">
                    <Point xmlns="http://www.opengis.net/gml">
                        <coord><X>45.3333333</X><Y>29.6666667</Y></coord>
                    </Point>
                    </adlgp:footprint>
                    <adlgp:footprint><gaz:CSquares>1204:495:363</gaz:CSquares></adlgp:footprint>
                    <adlgp:footprint><gaz:GeoHash>u8j1ghdtg3y4</gaz:GeoHash></adlgp:footprint>
                    <adlgp:footprint><gaz:WKT>POINT (45.3333333 29.6666667)</gaz:WKT></adlgp:footprint>
                </adlgp:footprints>
                <adlgp:classes>
                    <adlgp:class primary="true"
                                thesaurus="ADL Feature Type Thesaurus">streams</adlgp:class>
                    <adlgp:class primary="false"
                                thesaurus="Geonames Feature Type Thesaurus">H</adlgp:class>
                </adlgp:classes>
                <adlgp:relationships>
                    <adlgp:relationship relation="part of"
                        target-identifier="http://sws.geonames.org/798549/" target-name="România"/>
                    <adlgp:relationship relation="in country"
                        target-identifier="http://sws.geonames.org/798549/" target-name="România"/>
                </adlgp:relationships>
            </adlgp:gazetteer-standard-report>
        </sru:recordData>
    </sru:record>
</sru:records>
</sru:searchRetrieveResponse>
```

Figure 6.11: SRU “Danube” streams query response

6.4 Administration Manager Service

The Administration Manager is, as the name indicates, an external service responsible for managing, in this case, a set of other services that are designed for explicit purposes. Its main function is as a control center linked to the Administrative UI, from which an Administrator can check the services, invoke operations from them and other tasks.

Some services run semi-autonomously such as the Duplicate Detection services, which may request for human resolution of conflicts. Other services, such as the Ingester, require only the definition of some parameters (collection to import/ feature(s) to export) and then run autonomously, producing the required output (exported features) or confirmation of success (Ingestion successful).



Search Results

Total Records Found: 1

Records Returned by SRU:

Tag Name	Value
ID:	http://sws.geonames.org/791630/
Name	Danube River
Alternate Names	• Danube River • Dunai • Dunărea (ro) • Dunaj • Danubio (es) • Danubo (eo) • Dunav • Danubius (la) • Danube (en) • Dunay • Duna • Donau (de)
Country	RO
Point Coordinates	(45.3333333, 29.6666667)
Classifications	• **streams**(ADL Feature Type Thesaurus) • **H**(Geonames Feature Type Thesaurus)
Relations	• *part of* - **România** • *in country* - **România**

Figure 6.12: SRU "Danube" streams HTML query response

6.4.1 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.

At the moment, the service supports importing content from Geonames (by importing a RDF data dump) and from Geo-Net-PT 02[3] (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] `http://xldb.fc.ul.pt/wiki/Geo-Net-PT_02_in_English`

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[4] - a data aggregator that exposes its stored data via Open Archives Initiative - Protocol for Metadata Harvesting (OAI-PMH)[5], so as to delegate responsibility to REPOX for checking sources for updated data.

6.4.2 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 Chapter 3, 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.

[4] `http://repox.ist.utl.pt/`

[5] `http://www.openarchives.org/pmh/`

- The name comparisons are made through use of an implementation of the Jaro-Winkler similarity metric, from SecondString[6], 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.

$$\begin{aligned} d &= \arccos(\sin(lat1) * \sin(lat2) + \cos(lat1) * \cos(lat2) * \cos(long2 - long1)) * R \\ R &= \textit{Earth-Radius}, 6,371\text{km} \end{aligned} \tag{6.1}$$

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.

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

Chapter 7

Statistics and Conclusions

This chapter serves to present some statistical data for the Gazetteer System data stored, obtained by means of the Ingester and classified through the Duplicate Checker. Then we present the conclusions drawn from the obtained results and findings with the completion of this work. We also explain the completed work and its differences with the planned solution. Finally, we present future work that will be performed and also other options or lines of work to consider pursuing in future endeavors.

7.0.3 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[1]. This section the statistics regarding the two data-sources imported and statistics for the end result internal database.

Table 7.1 shows the statistics regarding the content from the Geonames RDF dump. Table 7.2 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 associated	5391
Countries - Least Features	NF - Norfolk Island (5)
Countries - Most Features	US - United States of America (2059852)
Feature Types - Total	618
Features - No Feature Type associated	5502
Features - Average p/Feature Type	12103
Features - Total	7479708
Features - Dataset size	10140 MB

Table 7.1: Geonames data statistics

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

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

Table 7.2: Geo-Net-PT 02 data statistics

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 7.4 shows the resulting database statistics.

Features - Number of Place Names	7026118
Features - Number of Feature Types	618
Features - Number of Footprints	7487512
Features - Relationships	15398856
Features - Relationship Types	4 (partOf, contains, adjacentTo, inCountry)
Features - Total	7684437
Number of DataSets used	2
Features - Size	12324 MB (not counting the original dataset tables)

Table 7.3: 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 graphics 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 in Chapter 3. 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 figure shows a summary of the results from that test. 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.

7.0.4 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

Features - Total for Batch	553552
Features - Found Duplicate Candidates	1164
Features - Duplicates with Distance Margin below 1Km	40

Table 7.4: Duplicate Detection data statistics

(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.

The result of these additional requirements was the creation of a basic structure of a minimal standard Gazetteer service compliant with ISO and OGC standards along with an “ecology” of services and interfaces, designed to enhance the gazetteer service and provide better access to its contents.

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 considerable 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. There was also

Overall, despite the differences, the main goal was set, as the service was accepted and positively reviewed.

7.1 Future Work

Because the final deadline for delivery for this project is set for September, there is still room for improving and fixing issues with the service and although not every issue can be dealt with appropriately in the remaining time-frame, these are the main points of interest I intend to tackle until the end of the project.

- Improve the Ingester so as to possibly work with the REPOX system, fetching all data provider content from one endpoint and also delegating responsibility of maintaing the provider’s data updated to REPOX:

- Expand the Duplicate Detector to include more combinations of techniques in order to improve precision;
- Attempt to get the data in a format fit for publishing in the LOD Cloud;

Bibliography

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

[2] ISO/IEC IS 19107:2003: Geographic Information - Spatial Schema. International Organization for Standardization, Geneva, Switzerland.

[3] ISO/IEC IS 19111:2007: Geographic Information - Spatial Referencing by Coordinates. International Organization for Standardization, Geneva, Switzerland.

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

[5] 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.

[6] R. B Allen. A query interface for an event gazetteer. In *Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries*, pages 72–73, 2004.

[7] ANSI/NISO. ANSI/NISO Z39.19 - 2005 guidelines for the construction, format, and management of monolingual control led vocabularies (NISO press). 2005.

[8] Tim Berners-Lee. *Weaving the Web: The Past, Present and Future of the World Wide Web by its Inventor*. Texere Publishing, 2000. ISBN-13: 978-1587990182.

[9] Tim Berners-Lee. Linked Data - Design Issues. Architectural note, 2006. http://www.w3.org/DesignIssues/LinkedData.html.

[10] Tor Bernhardsen. *Geographic information systems: an introduction.* John Wiley and Sons, Third edition, 2002. ISBN-13: 978-0471419686.

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

[12] 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.

[13] 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.

[14] W.W. Cohen, P. Ravikumar, and S.E. Fienberg. A comparison of string distance metrics for name-matching tasks. In *Proceedings of the IJCAI-2003 Workshop on Information Integration on the Web (IIWeb-03)*, pages 73–78. Citeseer, 2003.

[15] Jeff de la Beaujardiere. OGC Web Map Server implementation specification. *Open Geospatial Consortium*, 2006.

[16] L.R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

[17] J. Drucker and B. Nowviskie. Temporal modeling: conceptualization and visualization of temporal relations for humanities scholarship. 2005.

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

[19] J. Fitzke and R. Atkinson. OGC best practices document: Gazetteer Service-Application profile of the web feature service implementation specification-0.9. 3. *Open Geospatial Consortium*, 2006.

[20] A. J Flanagin and M. J Metzger. The credibility of volunteered geographic information. *GeoJournal*, 72(3):137–148, 2008.

[21] M. F Goodchild. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4):211–221, 2007.

[22] M. F Goodchild. Spatial accuracy 2.0. In *Proceedings of the Eighth International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*, volume 1, pages 1–7, 2008.

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

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

[25] P. Harpring. The limits of the world: Theoretical and practical issues in the construction of the Getty Thesaurus of Geographic Names. In *Proceedings of the 4th International Conference on Hypermedia and Interactivity in Museums (ICHIM'97)*, pages 237–251, Paris, France, 1997.

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

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

[28] Michael Hausenblas. Linked data applications - the genesis and the challenges of using linked data on the web. Technical report, Digital Enterprise Research Institute (DERI), 2009.

[29] D. Heise. Event structure analysis: A qualitative model of quantitative research. *Using Computers in Qualitative Research*, pages 136–163, 1991.

[30] 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.

[31] J. R Herring. OGC implementation specification for geographic information simple feature access part 1: Common architecture. *Open Geospatial Consortium*, 2006.

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

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

[34] 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.

[35] (ISO). I.O.f.S. guidelines for the establishment and development of monolingual thesauri. 1986.

[36] P. Jaccard. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, (37):547–579, 1901.

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

[38] M. Jensen. Visualizing complex semantic timelines. In *Proceedings of the ACH-ALLC Digital Humanities 2006*, Paris, France, July 2006.

[39] T. Joachims. Making large-scale svm learning practical. *Advances in Kernel Methods Support Vector Learning*, pages 169–184, 1999.

[40] 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.

[41] C. Keßler, K. Janowicz, and M. Bishr. An agenda for the next generation gazetteer: Geographic information contribution and retrieval. In *International Conference on Advances in Geographic Information Systems*, volume 2009, 2009.

[42] C. Keßler, P. Maué, J. T Heuer, and T. Bartoschek. Bottom-Up gazetteers: Learning from the implicit semantics of geotags. *Lecture Notes in Computer Science - GeoSpatial Semantics*, 5892/2009:83–102, 2009.

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

[44] H. Manguinhas, B. Martins, J. Borbinha, and W. Siabato. The DIGMAP Geo-Temporal web gazetteer service. *e-Perimetron*, 4(1):9–24, 2009.

[45] A. McCallum, K. Nigam, and L.H. 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.

[46] A.E. 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.

[47] 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.

[48] R. Mostern. Historical gazetteers: An experiential perspective, with examples from chinese history. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 41(1):39–46, 2008.

[49] R. Mostern and I. Johnson. From named place to naming event: creating gazetteers for history. *International Journal of Geographical Information Science*, 22(10):1091–1108, 2008.

[50] 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.

[51] 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.

[52] 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, October 2009.

[53] B. Robertson. A developer's guide to the resources of the historical event markup and linking project. 2001-2005.

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

[55] 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.

[56] 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.

[57] V. Sehgal, L. Getoor, and P.D. 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.

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

[59] 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.

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

[61] 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.

[62] P. Vretanos. OGC Web Feature Service implementation specification. *Open Geospatial Consortium*, pages 02–058, 2002.

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

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