e-Perimetron, Vol. 4, No. 1, 2009 [9-24] www.e-perimetron.org | ISSN 1790-3769

H. Manguinhas*, B. Martins*, J. Borbinha*, W. Siabato**

The DIGMAP Geo-Temporal Web Gazetteer Service

Keywords: Gazetteer; Geographic; Temporal; Integration; Digital Libraries; DIGMAP.

Summary
This paper presents the DIGMAP geo-temporal Web gazetteer service, a system providing access to names of places, historical periods, and associated geo-temporal information. Within the DIGMAP project, this gazetteer serves as the unified repository of geographic and temporal information, assisting in the recognition and disambiguation of geo-temporal expressions over text, as well as in resource searching and indexing. We describe the data integration methodology, the handling of temporal information and some of the applications that use the gazetteer. Initial evaluation results show that the proposed system can adequately support several tasks related to geo-temporal information extraction and retrieval.

Introduction

DIGMAP[1] stands for *Discovering our Past World with Digitized Historical Maps*, but it could stand also for digging on maps! The project addresses information retrieval (IR) methods specific for digital libraries of old maps, supporting the searching and browsing of resources according to temporal and geographical criteria [10]. DIGMAP builds on previous efforts related to the area of geographical information retrieval (GIR), for instance the Alexandria Digital Library (ADL) project [1], SPIRIT [2] and other studies [3][4].
GIR research started with the idea that geography provides a powerful searching and browsing mechanism, but there is a semantic gap between user requirements and the functionality supported by standard Geographical Information Systems (GIS). Traditional GIS allow access to geospatial information in a spatial way, using primitives such as points and polygons. However, there is little support for the use of place names. Although typical digital libraries and IR systems lack geospatial capabilities, they commonly use place names to describe the resources. GIR aims to add geographic coordinates into these previously non geo-referenced resources, spatially enabling them. In GIR, gazetteers are typically used to support the conversion of place names into geographical coordinates.
A problem often ignored in GIR research, and that naturally occurs in a service like DIGMAP, is that both modern and historical place names are simultaneously used. Many resources indexed in DIGMAP indeed relate to regions that no longer exist. Finding historical names, understanding to what regions on the globe these names refer to at different times, and understanding how these names relate to modern geography, presents many challenges to existing gazetteer services.
This paper describes the gazetteer system developed in the context of DIGMAP. Generally, this can be seen as a database of geographical features (e.g. countries, cities, rivers, etc.), with descriptive information about their names, locations, temporal coverage and associations. In addition, the gazetteer also includes histori-

* Instituto Superior Técnico - Department of Computer Science and Engineering. Av. Rovisco Pais, 1049-001 Lisboa, Portugal [hugo.manguinhas@ist.utl.pt] [bruno.g.martins@ist.utl.pt] [jlb@ist.utl.pt]
** Universidad Politécnica de Madrid – Laboratory of geographic information technologies (LatinGEO) Campus Sur UPM. Km. 7.5 Autovía de Valencia, 28031, Madrid, España [wsiabato@acm.org]

[1] www.digmap.eu

cal periods, with descriptive information about their names, time-spans and relations to the geographical concepts.

In DIGMAP, the gazetteer supports tasks such as the geo-parsing (i.e. associating the references to places and historical periods, occurring over the metadata records, to the corresponding time-spans and geospatial coordinates) and indexing of the resources. The gazetteer is available through an XML Web service interface, similar to the one proposed for the Alexandria Digital Library (ADL) gazetteer [1]. Adaptors were developed for outputting the results in other popular formats, such as KML, geoRSS or OWL. Particular emphasis was given to the performance of the service, through the introduction of a simpler and more flexible data model than the one used in ADL, as well as caching and indexing mechanisms.

Building gazetteers is a non-trivial task that involves the integrated usage of heterogeneous information sources. The complexity of the problem is inherently related to the complexity and dimensions of the data. A place may have more than one name and multiple relations to multiple other places, which may also change over time. Moreover, data coming from different sources varies in many dimensions. We follow an extraction, transformation and loading (ETL) methodology, typical of data warehousing systems, for the integration of multiple data sources. The central repository follows the general organization of concepts proposed for the ADL gazetteer, introducing minor changes related to the temporal domain. The considered data sources include public gazetteers with world coverage (e.g. the GeoNames[2] dataset), together with smaller gazetteers and bibliographic authority files.

This paper is organized as follows: Section 2 presents concepts and related works; Section 3 outlines the gazetteer service, describing its main features; Section 3 describes the data integration methodology; Section 4 presents applications of the gazetteer service within the context of DIGMAP; Section 5 presents evaluation results; finally, Section 6 presents some conclusions and a discussion on future work.

Concepts and related work

Gazetteers have a fundamental role in automating the usage of place names, by providing the means for translating them into unambiguous geospatial coordinates. Besides supporting place name lookups, gazetteers can also hold other useful information for GIR applications. Facts about places that are often found on gazetteers include place type information, demographics, topological relations and geospatial footprints.

Currently, gazetteer data exists in many independent and often dissimilar sources. Examples include:

- Gazetteers of official toponymic authorities.
- Local or special purpose gazetteers.
- Indexes accompanying published atlases.
- Place identifier tables accompanying GIS datasets.
- Place authority files used for cataloging and indexing.
- Historical printed gazetteers and encyclopedias.
- Online sources such as Wikipedia.

Appendix A lists popular gazetteers currently available on the Web. Despite the increasing popularity of such resources, there are also many documented problems. Most gazetteers were built for specific purposes and not designed to be interoperable or shareable. Gazetteer data can vary in many dimensions (e.g. scope, completeness, correctness, granularity, balance and richness [9]) and there is no standardization on the formats and service interfaces.

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

Integrating data from multiple gazetteers remains an important research challenge. Previous studies proposed to use extraction, transformation and loading (ETL) methods for integrating data from multiple sources into a unified repository (e.g. a relational database) [13][5]. Challenges are related to duplicate detection and fusion [26], and to the definition of mappings for different typing schemes [27].

The degree in which gazetteers include spatial and temporal information is also variable. Spatial data in gazetteers is usually confined to simple representations (i.e. centroid coordinates). Moreover, although places and the associated facts change over time, few gazetteer services model temporal ranges for the data. Some of the resources in Appendix include historical names, but few contain this information cross-walked with temporal periods.

A particularly noteworthy example regarding the use of temporal information in gazetteers is the ECAI Time Period Directory [21]. This is a metadata infrastructure similar in style to the ADL geographic gazetteer but for named time periods, linking them to geographic locations as well as to canonical time ranges. It builds on a content schema for describing named time periods and linking them to dates and locations, also providing a type list for categorizing periods (e.g. reigns, wars, revolutions, etc.). Another previous work regarding geo-temporal gazetteers was reported in [17], introducing an event gazetteer storing and presenting locations in time.

Of all the resources listed in Appendix, the GeoNames geographic database is perhaps the most widely accepted. Available for usage and download under a creative commons attribution license, it contains over 14.5 million geographic names for more than 6.6 million unique features, describing about 2.3 million populated places and 8 million alternate names. GeoNames features have a unique identifier, a name, alternative names (e.g. in different languages), part-of relations to administrative divisions and geo-spatial coordinates. All features are categorized into one out of 9 classes and further subcategorized into one of 645 codes. GeoNames integrates data from various sources, mainly the Geonet Names Server (GNS) gazetteer of the National Geospatial Intelligence Agency (NGA), the Geographic Names Information System (GNIS) gazetteer of the U.S. Geographic Survey, the GTOPO30 digital elevation model for the world developed by United States Geological Survey (USGS) and information from Wikipedia. The most important limitations of GeoNames relate to the lack of historical place information and to the simple representation of spatial footprints as centroid coordinates.

The Getty Thesaurus of Geographic Names (TGN) is another well-known gazetteer service [25]. However, unlike GeoNames, usage of TGN data requires a private license. The TGN was compiled from different sources and contains about 1 million places around the globe, including both political entities (e.g. nations) and physical features (e.g. rivers). The focus of TGN records are places and each one has a unique ID. Linked to place records are names (common, historical and spelled in different languages), the place's hierarchical ancestor, other semantic relationships (e.g. equivalent and associative), geospatial coordinates, annotations, data-sources, and place types (e.g. inhabited place, state capital). There may be multiple hierarchical ancestors associated with each place, making the TGN poly-hierarchical. The dates associated with place names are expressed by two years delimiting a span of time. However, many names lack this information. Time spans are available in varying levels of specificity and certainty.

The Alexandria Digital Library (ADL) project addressed the development of gazetteer and thesaurus protocols to support search and retrieval over distributed resources [1]. This was one of the pioneering efforts in defining the basic elements of a content standard for gazetteer data. The ADL gazetteer content standard defines the core elements of named places (and their history), their spatial location (in various representations), classification (according to referenced typing schemes), and metadata properties (e.g. source attribu-

tion). The DIGMAP gazetteer service generally follows this model, therefore will be given further details in the remaining sections of this paper. In terms of the actual data, the ADL gazetteer combines the U.S. place names from GNIS and the non-U.S. place names from GNS, as well as other gazetteer datasets. An implementation of the ADL gazetteer service was also released as open source, although usage of the complete dataset requires a private license. From our initial experiments, there were some performance issues with the open source implementation. We therefore made a new implementation, using a simpler data model together with efficient caching and indexing strategies.

The EDINA GeoXWalk gazetteer [16] is a middleware service implementing a digital gazetteer for the UK academic community (i.e. a gazetteer of geographical features within Great Britain built predominantly from Ordnance Survey data). The rationale behind the project was to support geo-parsing and enhanced geospatial searching, and to provide reference services for spatial searching within the existing academic network. To the best of our knowledge, work within the GeoXWalk project did not address the temporal domain.

The Open Geographical Consortium (OGC) proposed a gazetteer service [14] based on a re-factored ISO-19112 content model published through a Web Feature Service (WFS). There are many similarities between OGC's proposal and the ADL gazetteer service. Implementations of the OGC gazetteer model are nonetheless scarce and, to the best of our knowledge, there is not a single one addressing issues related to the temporal domain.

In this work, we introduce a Web gazetteer service that stores and presents *locations in time and time periods in space*, essentially refining ideas from the ADL gazetteer and the ECAI time period directory.

The DIGMAP gazetteer

The DIGMAP gazetteer is an information system responsible for managing geographic and temporal information. The gazetteer acts as a middleware service (machine-to-machine) within the DIGMAP architecture, supporting for other services requiring place and time period data (e.g. a geo-parser service for processing textual documents).

Data in the DIGMAP Gazetteer

The data within the DIGMAP gazetteer is defined using OWL, offering a formal way for the representation of information, while adding a greater level of expressivity. A new ontology was developed using the description logics part of OWL. This ontology defines a feature class considering six core properties: identification (internal and source identifications), names, spatial footprints, temporal coverages, feature types, associations to other features, and metadata (e.g. demographics). 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, and there may be associations between the two for cross-walking between temporal and geographical domains. Geographic features must always be associated with names. However, for temporal features, the specification of time spans alone is also allowed. Figure 1 provides an illustration.

In terms of the typing scheme chosen to categorize the features, we use a combination of the ADL Feature Type Thesaurus (FTT) with the classification scheme from the ECAI Time Period Directory. This schema is hereby referred to as the DIGMAP Feature Type Ontology (DFTO). All gazetteer features are always associated with a feature type in DFTO. In practice, the DFTO is also an OWL ontology defining classifica-

tion terms and relationships among them. Besides defining primary typing schema, the DFTO contains mappings between the primary types and other classification schemas (e.g. the GeoNames classification schema) in order to facilitate data integration from external sources. A large set of such mappings was manually defined and included in the OWL ontology.

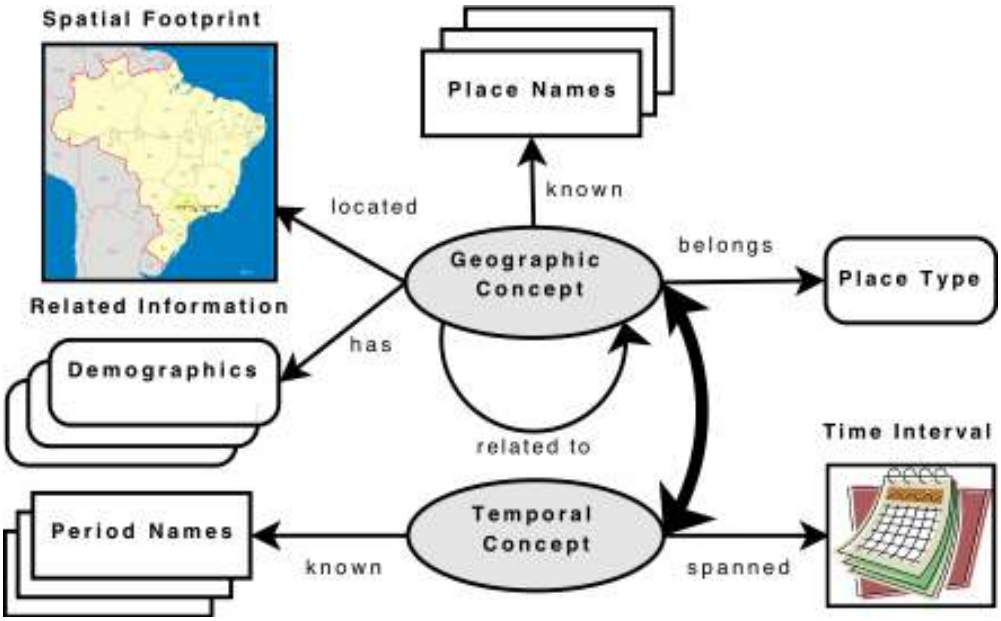

Figure 1. Core elements of gazetteer data.

In terms of the semantic relations that can be defined for features, *having* relations are used to define the associations: between features and DFTO types; between features and metadata properties; between geographic features and spatial footprints; and between temporal footprints and time spans. Between temporal and geographic features it is possible to have in-*context-of* relationships. Between each property of the geographic features (e.g. the names, spatial footprints or metadata elements) and temporal features, *in-context-of* relations are also possible. Finally, between the features themselves, the *part-of*, *contains*, *equivalent* and *adjacent* relationships are enabled. More relations are intrinsically encoded in the time spans and the geospatial footprints (e.g. a distance and overlap among geographical footprints, or a temporal ordering for the time spans).

Geospatial footprints are defined as both GML strings representing points, bounding boxes or polygons, and C-Squares strings [6] obtained from the GML geometries.

Storage and Encoding of Gazetteer Data

The data in the DIGMAP Gazetteer is stored in a relational database as records encoded in XML (each OWL 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.

To endorse requests in different encoding formats, the DIGMAP gazetteer uses XML stylesheets (XSLT) to transform the data. The system was designed to support one master database holding the records encoded in the internal gazetteer format, and several cache databases to hold the other encoding formats. The transla-

tions are done only the first time they are requested, and stored in the cache database for further reuse, thus improving the performance of the system.

Machine access to gazetteer content

As previously stated, the gazetteer follows the service protocol and query language defined in the ADL project with some minor modifications (e.g. support query filters for name similarity and geo-temporal restrictions). Besides providing results in the ADL gazetteer standard, other popular formats are also supported, e.g. KML[3], geoRSS[4], or the XML format defined by the OGC for gazetteer Web services (WFS-G). Regarding the query format, both the ADL and WFS-G formats are supported. XSLTs are, once again, used to transform queries in the WFS-G format into ADL queries.

Query Language

The DIGMAP gazetteer follows the query language specification provided for the ADL gazetteer. The ADL query model already supported complex queries with textual, spatial or typing restrictions, as well as Boolean operations for their combination. Some differences were introduced for the DIGMAP gazetteer:

- Users can request temporal features by name or time span using temporal filters (**equal**, **included in**, **starting before**, **ending before**, **starting after**, **ending after**, and **including** the time stamp). This is similar to querying geographical features by their names or spatial footprints.
- Users can use a complex set of spatial filters (**equal, maximum distance, contained, containing, overlapping and outside**) when requesting geographical features. This is based on OGC's Filter Encoding specification [15]. For providing spatial footprints, either GML or C-Squares strings can be given.
- Users can search for **similar** names, either in terms of character differences or phonetics, a part from the existing name searches (e.g. containing **all the words**, **any words**, **phrase**, **name equals**, and **regular expressions**). The similar names option corresponds to a combination of the Jaro-Winkler similarity measure [19] with the double metaphone phonetic similarity algorithm [18].
- Users can crosswalk geography and time, by querying temporal features related to a given geographic feature or a spatial footprint, and by querying geographic features having properties related to a given temporal feature or time span.
- Users can search for features using filters for the feature data information (**equal**, **greater**, **smaller** than a query value).

Implementation

In terms of implementation, the gazetteer service was designed to implement the above query options with high efficiency. For increased performance, the gazetteer builds indexes in a relational database for each query option. This is done using a relational database (i.e. Apache Derby[5]) together with specific APIs for evaluating spatial properties, namely GeoTools[6] and an open-source geomatics engine called Java Topol-

[3] http://code.google.com/apis/kml/
[4] http://www.georss.org/
[5] http://db.apache.org/derby/
[6] http://geotools.codehaus.org/

ogy Suite[7] (JTS). JTS natively supports the GML format and it also implements several multi-dimensional indexing strategies [28]. Although any relational database could in principle be used, Apache Derby has the advantage of providing a deep integration with Java. It is possible to call Java methods (e.g. JTS methods or string similarity functions) from SQL queries, making it easier to implement complex query filters.
When queried, the indexes return the identifications for the features that were matched, which are then used to retrieve the actual records from the storage database and build the response query.

Figure 2. The DIGMAP Gazetteer user interface.

Human Access to the Data

Besides the XML Web service, a simple user interface was also developed to support data insertion and retrieval by human users (see Figure 2). Data insertion and updating is made through a form that, when submitted, generates an XML file that is send to the service interface. Having human users inserting information into the gazetteer is therefore no different that integrating information from an external source. As for retrieval, it allows expert users to introduce queries using the XML protocol, along with simple keyword-based queries for casual users. The results can be seen as an HTML page containing both a textual report and a set of markers presented over a dynamic map. This is also illustrated in Figure 2.

Integration of gazetteer data

As previously stated, existing sources of gazetteer data vary in many dimensions. Some impose restrictions to the usage of the data and others are not even available in a structured digital form. A large dataset with a worldwide coverage was chosen to populate the DIGMAP gazetteer, complemented though with smaller but more specialized sources. Currently, the DIGMAP gazetteer integrates data from the following data sources:

- GeoNames data available for download as a text file.

[7] http://www.vividsolutions.com/jts/

- The GeoNetPT OWL ontology, including modern Portuguese place names, demographics information and detailed spatial footprints.
- Place names at authority records from the Estonian National Library, available in the XML MADS format.
- Time period names from the ECAI time period directory.
- Information from Wikipedia concerning alternative names of places and historical periods, extracted by hand.

The general procedure for integrating data from the previous sources follows the typical ETL approach of data warehouse systems. It starts with the creation of wrappers from the original format into the internal format (i.e. extraction and transformation). These XML files are then integrated into the gazetteer database (i.e. loading). Figure 3 provides an illustration.

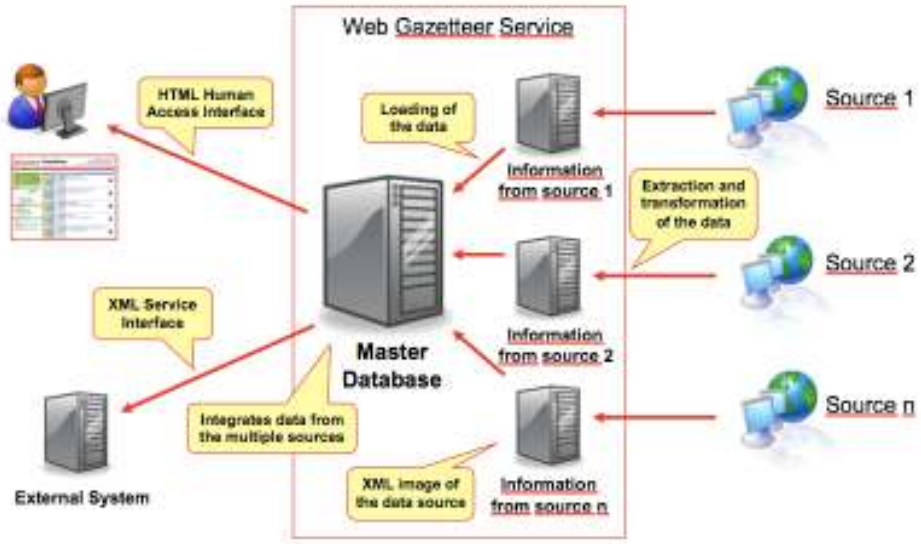

Figure 3. Integrating data into the gazetteer.

When integrating data from these multiple sources, the problem of variable typing schemes was obviated by having two-way mapping associations between the types defined at the original sources and the types defined at the DFTO. Geographic features are this way always defined according to a consistent coding convention, but without ever loosing the original information. Users can use the service to query for place names with basis on the geographic types defined in the original sources.

A more challenging data integration problem relates to checking if two pieces of gazetteer data are about the same feature (i.e. the same place). This presents a difficult challenge because no single piece of data about a feature is unique. In the geographical case, the same name can be linked to different places and a place can have multiple names, either because of different languages, variant spellings or changes through time. Spatial footprints also come in different forms (e.g. points or polygons) and at different resolutions. Finally, there can be variations according to different time periods.

Presently, there is no duplicate elimination for geographical features. Some tests have been made, looking at three main feature properties (names, feature types, and spatial distance between features) with interesting results. For measuring name similarity, we plan to use the Jaro-Winkler [19] measure, adapted to ignore

diacritics, in combination with the double metaphone phonetic similarity algorithm [18]. Spatial similarity will be based on distance and overlapping metrics, provided through the JTS/GeoTools APIs.
Another problem in populating the gazetteer with data from multiple sources relates to complementing the geo-temporal relationships that are defined in the datasets. Sources like the GeoNetPT OWL ontology already capture many conceptual relations among geographical features (i.e. a part-of hierarchy, equivalence and adjacency) but other datasets are not so rich. There are many associations that can be inferred from the data, either through simple inference procedures or by geographic computations (e.g. using distance or area). When inserting data into the gazetteer, reasoning mechanisms are used to add additional geo-temporal relationships.
The most typical form of reasoning rules for composition of spatial relationships is the so called *triangulae knowledge*, stating that $\forall x,y,z: rel1(x,y) \wedge rel2(y,c) \Rightarrow rel3(x,c)$. The following rules are a subset of the ones that were considered for our gazetteer, using mechanisms similar in style to the *triangulae knowledge* rule:

```
∀x,y : partOf(x,y) ⇒ contained(y,x)
∀x,y : adjacent(x,y) ⇒ adjacent(y,x)
∀x,y : equivalent(x,y) ⇒ equivalent(y,x)
∀x,y : spatialInside(x,y) ⇒ partOf(x,y)
∀x,y : spatialCoveredBy(x,y) ⇒ contained(x,y)
∀x,y,z : partOf(x,y) ∧ partOf(y,c) ⇒ partOf(x,c)
∀x,y,z : equivalent(x,y) ∧ equivalent(y,c) ⇒ equivalent(x,c)
∀x,y,z : spatialInside(x,z) ∧ (spatialEqual(z,y) ∨ spatialInside(z,y) ∨ spa-
tialCoveredBy(z,y)) ⇒ part-of(x,y)
```

By interleaving forward and backward reasoning, new facts can be derived. This procedure is done offline, whenever a new dataset is integrated into the gazetteer.

Challenges related to the temporal domain

Just as locations are commonly referred to by place names as opposed to spatial footprints, temporal periods are also commonly referred to by names such as *Renaissance* or *Napoleonic Wars*, although periods could also be unambiguously specified through the use of dates.
In the DIGMAP gazetteer, the storage and access to names of historical periods has been designed to mirror the treatment of the geographic concepts. Each period can be described by several names and has an associated time span.
For now, the time periods that are defined in the gazetteer were mainly extracted from the ECAI Time Period Directory, and this information was then complemented by hand with translations to other languages and other temporal periods described over Wikipedia pages[8].
One of the main motivations for the inclusion of temporal information in the gazetteer relates to the fact that geographic regions change over time. They can be split (e.g. former Czechoslovakia), merged together (e.g. former East and West Germany) or have their names changed (e.g. Zaire changed its name to Congo). As a result, queries to the gazetteer may contain names that do not exist anymore, or that refer to regions that are different from those that are currently referred to. To represent these aspects, specific relations in the gazetteer ontology are used, which cross-walk time and geography. Besides the definition of time peri-

[8] http://en.wikipedia.org/wiki/List_of_time_periods

ods (which can be very useful in itself for tasks such as the recognition of period names in text), the properties of geographical features can also be associated with specific time periods.
The representation of these relationships is relatively simple, and is already described in previous sections. However, an important concern is the lack of historic data in most of the gazetteer datasets that are currently available, and much less information regarding the temporal extents that are associated to particular geographical features. The identification of spatial boundaries and feature types for historical data is in itself quite challenging, as the methods of modern cartography often do not apply.
In the DIGMAP gazetteer, we plan to address these very difficult issues through the following strategies:

- Allow human users to insert and edit the information that is stored in the gazetteer. Since the ontology is sufficiently rich to support geo-temporal associations, human users can in time provide rich data to the service. This is the principle behind Web sites such as Wikipedia. Our gazetteer already contains many new associations and corrections to the data, which were introduced by people currently involved in the DIGMAP project.
- Automatically explore information sources such as metadata records in digital library catalogues. These records often contain indexing information using modern place names, and descriptive information containing the equivalent historical names. The same records can also contain indexing information relating to time. Cross-linking the time information with the historical place names can provide estimates for the time-spans associated with the usage of some place names.

The study of these techniques is currently ongoing work.

Applications of the gazetteer service

Initial requirements analysis for the DIGMAP project raised the need for a geo-parser, i.e. a software service that can take textual resources (e.g. metadata records in library catalogues) possibly containing names for places and historical periods, automatically identify the occurrence of such references and finally assign the resources to encompassing geo-temporal scopes. The gazetteer service offers the necessary support, providing mechanisms for matching references in the text against gazetteer entries.
The DIGMAP geo-parser works as follows: standard information extraction techniques are first used to find relevant references in the text. Following the identification of possible geo-temporal references, each of them is disambiguated into the corresponding gazetteer feature(s). The document is finally assigned to an encompassing geo-temporal scope, determined with basis on the most general gazetteer features that combine the references made in the text. This is detailed in a separate publication [22].

Evaluation

This section describes initial evaluation experiments performed with the proposed gazetteer service, starting with a summary on the results obtained with a geo-parser system that uses the gazetteer, and continuing with the presentation of statistical characterization results.

A Geo-Parser Service Using the Gazetteer

A separate publication already presented evaluation results on a geo-parsing system that used the DIGMAP gazetteer for associating geo-temporal references in the text into unambiguous identifiers, as well as for assigning documents to encompassing geo-temporal scopes [22]. In terms of accuracy, the results were

much better for geographic than for temporal references. For instance, over 70 percent of the documents used in our experiments could be assigned to geographic scopes with an error of less than 100 Kilometers. Temporal references were only recognized in less than 10 percent of the documents. This can indicate that the gazetteer is still lacking in names for historical periods. In terms of performance, the geo-parser service did not encounter major problems in the usage of the gazetteer service.

Statistic	Value	Comment
Number of places	7.034.538	approx. 1/3 correspond to populated places
Number of place names	15.026.983	
Number of place types	210	Preferred terms in the ADL-FTT
Places with specific place type	6.900.377	
Number of historical periods	1.989	ECAI Time Period Directory + Wikipedia
Places with spatial footprints	66.211.38	Mostly centroids, a few bounding boxes
Number of relationship types	5	
Number of places with relations	431.397	Mostly from GeoNETPT
Number of place relations	866.019	Mostly *part-of* and *contains*
Number of time/place relations	1.989	

Table 1. Statistical characterization of the DIGMAP gazetteer

Statistical Characterization

This section presents a statistical characterization of the gazetteer dataset. The values presented in Table 1 reflect the gazetteer content after integration of the four data sources (GeoNames, GeoNetPT, Authority records, ECAI time period directory combined with Wikipedia). About one third of the geographic features that were gathered correspond to populated places (e.g. cities, districts, villages). Each geo-feature has an average of two names, but only one out of sixteen geo-features contains relations to other places. This is because the GeoNames database dump files that were used as the source of information do not contain the relationships (richer data can hopefully be collected from GeoNames using their semantic web portal[9]). Most of these relations come from the GeoNetPT data source, corresponding to *part-of* and *contains* relations.
In respect to temporal features, the number of historical periods and corresponding relations to places is still very small, due to the lack in data sources for this kind of information.

Conclusions and future work

Interest in geographic information technologies, particularly those related to places and place names, has grown significantly over the past few years. The powerful simplicity of application such as Google Earth fueled a wealth of geo-related activities and many on-going projects are also addressing the usage of place name information to build large interoperable spatial data infrastructures (SDIs). An essential component of any spatial-data infrastructure, supporting the retrieval of resources that are geo-referenced through the use of place

[9] http://sws.geonames.org

names, is a gazetteer service. Within the SDI, a gazetteer should model the terminology and associated structure of the geographic space.

Having gazetteers capable of cross-walking geographic and temporal information can be extremely useful for many applications. An interesting example is the linking of online library catalogs to information about places. Previous works have already concluded that scholars search in three major categories, namely biography (persons), chronology (periods) and geography (places) [24]. Efficient methods for exploring the geo-temporal domain can transform information searching throughout libraries and the Internet.

This paper presented the DIGMAP geo-temporal gazetteer service, a system integrating data from multiple sources and providing access to names of places, historical periods, and the associated geo-temporal information. This service is novel, in the sense that it stores and presents *locations in time and time periods in space*, refining ideas from previous works such as the ADL gazetteer and the ECAI time period directory. Within the DIGMAP project, the gazetteer serves as the unified repository of geographic and temporal information, assisting in the recognition and disambiguation of geo-temporal expressions over text, as well as in resource searching and indexing. Evaluation experiments attested for the adequacy of the proposed service interface, as well as to the usefulness of the gazetteer service in other DIGMAP tasks [22].

For future work, we will focus on the problem of duplicate detection and fusion. We will also perform additional evaluation experiments, particularly focusing on measuring the performance of the Web service for different types of queries. Finally, advanced techniques for enriching the data will also be experimented, for instance using semi-supervised learning methods for extracting information from the Web [8], using map data to enrich the gazetteer with more detailed spatial information [7], or using Voronoi polygons derived from centroid coordinates [11].

Acknowledgements

This research was partially funded by the eContentplus Programme of the European Community, under the contract ECP-2005-CULT-038042 (DIGMAP project).

References

[1] L. Hill and Q. Zheng 1999. Indirect geospatial referencing through place names in the digital library: Alexandria Digital Library experience with developing and implementing gazetteers. Proceedings of the American Society for Information Science Annual Meeting

[2] C.B. Jones, A.I. Abdelmoty, D. Finch, G. Fu and S. Vaid 2004 The SPIRIT Spatial Search Engine: Architecture, Ontologies and Spatial Indexing. Proceedings of the 3rd International Conference on Geographic Information Science

[3] Reid, J. 2003 geoXwalk - A gazetteer server and service for UK Academia. Proceedings of the 7th International Conference on GeoComputation

[4] I. Johnson 2004. Putting Time on the Map: Using TimeMap for Map Animation and Web Delivery, GeoInformatics

[5] M. Chaves and M. J. Silva and B. Martins 2005 A Geographic Knowledge Base for Semantic Web Applications, Proceedings of the 20th Brazilian Symposium on Databases

[6] T. Rees 2003 C-Squares, a New Spatial Indexing System and its Applicability to the Description of Oceanographic Datasets, Oceanography, 16(1)

[7] A. Khotanzad and E. Zink 2003 Contour line and geographic feature extraction from USGS color topographical paper maps, IEEE Transactions on Pattern Analysis and Machine Intelligence, 25(1)

[8] O. Uryupina 2003 Semi-supervised learning of geographical gazetteers from the Internet. Proceedings of the HTL/NAACL-03 Workshop on The Analysis of Geographic References

[9] J. L. Leidner 2004 Towards a reference corpus for automatic toponym resolution evaluation. Proceedings of the 1st Workshop on Geographic Information Retrieval

[10] B. Martins, J. Borbinha, G. Pedrosa, J. Gil, and N. Freire 2007 Geographically-aware information retrieval for collections of digitized historical maps. Proceedings of the 4th Workshop on Geographical Information Retrieval

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

[12] W.-F. Riekert 2002 Automated Retrieval of Information in the Internet by Using Thesauri and Gazetteers as Knowledge Sources. Journal of Universal Computer Science, 8(6)

[13] A. Axelrod 2003 On building a high performance gazetteer database. Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references

[14] J. A. Fitzke 2006 OGC Best Practices Document: Gazetteer Service - Application profile of the Web Feature Service Implementation Specification - 0.9.3. Open Geospatial Consortium

[15] A. V. Panagiotis 2005 OpenGIS Filter Encoding Implementation Specification. Open Geospatial Consortium

[16] J. Reid 2003 geoXwalk -- a gazetteer server and service for UK academia. Proceedings of the 7th International Conference on GeoComputation

[17] R. B. Allen 2004 A query interface for an event gazetteer, Proceedings of the 4th ACM/IEEE-CS joint conference on digital libraries

[18] L. Philips 2000 The Double Metaphone Search Algorithm, C/C++ Users Journal

[19] W. E. Winkler 2006 Overview of Record Linkage and Current Research Directions. Research Report Series

[20] Chen, Ya-ning Arthur, Shu-juin Sophy Chen and Wei-long Ueng 2002. Digital Gazetteer Service in Context of Chinese Culture, PNC

[21] V. Petras, R. R Larson and M. Buckland 2006. Time period directories: a metadata infrastructure for placing events in temporal and geographic context. Proceedings of the 6th ACM/IEEE-CS joint conference on digital libraries

[22] B. Martins, H. Manguinhas and J. Borbinha 2008 Extracting and exploring geo-temporal semantics of textual resources (to appear)

[23] H. Uitermark et al. 1999 Ontology-Based Geographic Dataset Integration. Proceedings of the International Workshop on Spatio-Temporal Database Management

[24] M. J. Bates and D. N. Wilde 1993 An analysis of search terminology used by humanities scholars: the Getty Online Searching Project Report Number 1. Library Quarterly, 63(1)

[25] P. Harpring 1997 The limits of the world: Theoretical and practical issues in the construction of the Getty Thesaurus of Geographic Names. Proceedings of the 4th International Conference on Hypermedia and Interactivity in Museums, Archives and Museum Informatics

[26] J. Hastings and L. Hill 2002 Treatment of Duplicates in the Alexandria Digital Library Gazetteer. Proceedings of GeoScience 2002

[27] Daniela F. Brauner, Marco A. Casanova and Ruy L. Milidiú 2006 Towards Gazetteer Integration Through an Instance-based Thesauri Mapping Approach. Proceedings of the 8th Brazilian Symposium on GeoInformatics

[28] V. Gaede and O. Gunther, 1997, Multidimensional Access Methods, ACM Computing Survey.

Appendix A

The table bellow describes some of the gazetteer services currently available in the Web.

Gazetteer Name	Scope	Temporal	Spatial data	Concepts	Names
Alexandria Digital Library Gazetteer	World	Limited	Points or MBRs	4.334.146	
GeoNames	World	Limited	Points	6.603.141	14.592.444
GeoNetPT	Portugal	No	Points or MBRs	431.397	434.539
U.S. Gazetteer	U.S.A.	Limited	Points	92.689	
Gazetteer for Scotland	Scotland	Yes	Very limited	13.471	
Global Gazetteer	World	No	Points		
National Gazetteer of Australia	Australia	No	Points	322.328	
Gazetteer of British Place Names	Britain	Yes	National grid code		50.000
Virginia Gazetteer	Virginia	No	USGS quadrangle	51.000	
Imperial Gazetteer of India	India	Yes	No		
The Fuzzy Gazetteer	World	No	Points		7.205.433
Getty Thesaurus of Geographic Names	World	Yes	Points or MBRs	912.000	1.100.000
W. Hazlit's Classical Gazetteer	World	Yes	No	5.000	
Maplandia Gazetteer	World	No	Polygons	166.000	
Geographical Names of Canada	Canada	No	Points	500.000	
Gazetteer of Tibet and the Himalayas	Tibetan regions in China	Yes	Points		
Old World Trade Routes Gazetteer	Eurasia + Africa	Yes	Points	3.130	12.500
National Gazetteer of Wales	Wales	No	National grid	6.000	
Bulgarian Antarctic Gazetteer	Antarctica	No	Points	97	97
US HomeTownLocator Gazetteer	U.S.A.	No	Points		
Markets/Fairs in England and Wales	England/Wales	Yes	No	2.400	
Orbis Latinus Gazetteer	World	Limited		16.352	
BSC Latin Place Names File	World	Limited		433	
Gazetteer of names of printing towns	World	Limited			
CERL Thesaurus	World	Limited			
Place Names Data at EKI	Estonia + more	Limited			
Roman place names	World	Limited			
Spanish gazetteer	Spain	No			
World Gazetteer	World		Points		
The Columbia gazetteer of the world	World	Limited	Points and some features		165.000
The Columbia gazetteer of North America	North America		Points/Data		50.000+
Geoscience Australia Place Name	Australia		Points		310.000+
Gazetteer of the Roman world	Roman Empire		Points/Data		
The ancient library	World	Yes	Points		15.000 +
Ordnance gazetteer of Scotland	Scotland		Points/Data		

Gazetteer of Slovakia	Slovakia		Points/Data		
U.S. Board on Geographic Names	U.S.A.		Points		
A gazetteer of Vermont places: real and imagined	Vermont		Points		
East and west Prussia gazetteer	Prussia		Points		
Newfoundland and Labrador Place Name gazetteer	Labrador (Canada)		Points		
NGA GEOnet Names Server	World		Points		7.000.000
Canadian Geographical Names Service	Canada	No	Points	350.000	
National Association of Counties	U.S.A.		Points/Data		
GeoNative	Athens		Points		
The Swedish gazetteer	Swedish		Complete features		57.000 +
Composite gazetteer of Antarctica	Antarctica		Points		36.000 +
CGDI gazetteer interface	Canada		No		47.000
A low-latitude Antarctic gazetteer	Antarctica Ext. References		Points		700 +
Old Hampshire gazetteer	Hampshire		Points		
Index Mundi	World				
Probert Encyclopaedia	World				70.000 +
Radix – 1882 gazetteer of Hungary	Hungary	Yes	Points		1.000.000
earthsearch.net	World		Points		7.400.00 +
German Space Operations Center gazetteer	World		Points		2.000.000 +
UK & Ireland gazetteers - GENUKI	U.K.		Points		
NYS gazetteer & GeoData Collection	New York State	No	Points		38.000
PlaceNames – South Australian State Gazetteer	South Australian State		Points		
Worldwide gazetteer	World	No	No		