A Geographic Knowledge Base for Semantic Web Applications

Marcirio Silveira Chaves[1], Mário J. Silva[1] and Bruno Martins[1]

[1]Departamento de Informática
Faculdade de Ciências - Universidade de Lisboa
1749-016 Lisboa, Portugal

Abstract. *This paper introduces GKB, a repository based on a domain independent meta-model for integrating geographic knowledge collected from multiple sources. We present the architecture, the repository design and the data cleaning and knowledge integration processes. We also describe the rules developed to add new knowledge to GKB. GKB includes tools for generating ontologies, which are being used by multiple semantic web applications. To illustrate how it is being used, we present some of the applications that interact with the repository or load ontologies created with GKB.*

Key-words*:* Information Integration, Knowledge Management, Ontology

1. Introduction

Web mining applications are receiving increasing attention both from academia [Etzioni et al. 2004, Sheth et al. 2004] and industry [Gruhl et al. 2004, Alexa 2005]. There is also an increasing interest on the analysis of web resources, focusing on their geographical context.

To support geographic semantics-aware web mining applications, we developed Geographic Knowledge Base (GKB). GKB integrates data and knowledge from multiple sources under a common schema, and doubles as an environment for deriving knowledge and generating ontologies from the available information. GKB maintains geographical information describing both geo-administrative and geo-physical entities and also the geographical attributes of network resources, such as web sites and internet domains.

The main contribution of this paper is an approach for creating and maintaining ontologies of geographic names for Semantic Web applications. We developed methods that support both the vertical and horizontal semantic integration of geographic data collected from very heterogeneous information sources. We believe that the proposed methods are general and their application to other ontologies of geographic names could be easily replicated. Many Web sites are multi-lingual and some text analysis applications require the identification of geographic locations in other languages than their local language. In GKB, we can specify alternative names for geographic entities and associate them to different languages.

Previous works created and used data and KBs for geographic information retrieval (IR) and geographic named entities recognition: Manov et al. developed an ontology as an alternative to flat structures of gazetteer lists [Manov et al. 2003]. Irie and Sundheim built an integrated geospatial database of place names information from four distinct gazetteers [Irie and Sundheim 2004]. In GKB, instead of using gazetteers as the single data source, we handle data from a diversity of information sources, from administrative authorities to information extraction tools. As Alani et al., our repository is based

on a generic meta-model, implemented as a relational database [Alani et al. 2003]. From the information gathered in this database, we generate ontologies to Semantic Web applications.

The Getty Thesaurus of Geographic Names (TGN) is a structured vocabulary including names and associated information about both current and historical places around the globe (`http://www.getty.edu/research/conducting_research/vocabularies/tgn/`). The focus of TGN records are places, each identified by a unique numeric ID. Linked to the record for the place are names (historical names, common alternative names and names in different languages), 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). There may be multiple broader contexts, making the TGN polyhierarchical. In addition to the hierarchical relationships, the TGN has equivalent and associative relationships. The structure and data of GKB is similar to TGN. However, we focus on Portuguese data and our resource is public and freely available.

We represent knowledge in Description Logics (DL) [Baader et al. 2003], the formalism adopted by the Semantic Web for this purpose. In addition to a common repository, GKB includes 2 sets of tools: *converters* load data from various sources, while performing some amount of data normalization to maintain a single unified view of all the information; *generators* create ontologies, following the OWL (Web Ontology Language) standard [McGuinness and van Harmelen 2004].

We have developed two GKB instances: the first is loaded with detailed information about the main geographic names of Portugal; the second, holds information about the main regions around the world in four different languages. It supports multi-language data and it allows loading of data about the main countries, cities and places around the world among other.

The remainder of this paper is organized as follows: the next Section presents the conceptual design of GKB and describes its information model. Section 3 discusses the data cleaning and knowledge integration methods used in GKB. We describe GKB as an ontology and some statistics of the data in our repository in Section 4. Section 5 describes some of the applications that are using it. Finally, Section 6 presents our final conclusions and some directions for future work.

2. Conceptual Design of GKB

Figure 1 gives an overview of GKB. Data is organized in information domains, each representing a set of related geographic features. There are presently three domains defined in GKB: geo-administrative, geo-physical and network. The information in each domain is structured identically, as they all implement a common meta-model.

GKB supports the definition of ontological relationships among the features of each domain. For instance, for the geographic domain, GKB essentially provides a hierarchical naming scheme with transitive "sub region of" and name alias capabilities. Tudhope et. al. listed the three main thesaurus relationships: i) equivalence (equivalent terms), ii) hierarchical (broader and narrower terms), and iii) associative (related terms) [Tudhope et al. 2001]. GKB provides these three types of relationships among

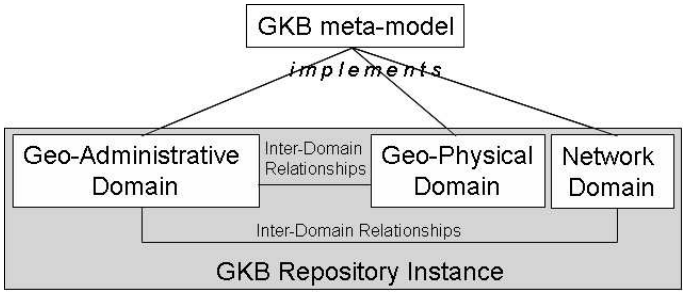

Figure 1. GKB Information Architecture

geographic features, specializing the associative relationship into generically associated and geographical adjacency. In addition, GKB also supports inter-domain relationships, which are associations between entities from different information domains. For example, we represent the geographic scope of a web resource as a relationship between a web site (a network domain entity) and a geographic region (a geographic domain entity).

GKB collects data from several classes of information sources. For the geographic domain, we have the following classes:

Administrative: contains data concerning demographics and administrative information, such as the territorial division. For Portugal, this kind of information comes from both Instituto Nacional de Estatística (INE) and the wikipedia on-line encyclopedia (at `http://en.wikipedia.org`). The latter also provides names of the main regions around the world for GKB instanced with world data. In addition ANMP (*Associação Nacional de Municípios Portugueses*) provides us the adjacency relationships among the districts and municipalities.

Postal: includes information used to identify addresses. For Portugal, we use the Portuguese Post Office (CTT) database, which publishes a database of postal codes. From this database, we get, for each postal code, the associated administrative area.

Gazetteer: provides information about the main cities, towns, and regions and their geographic coordinates. For Portugal, we use the calle web site data (at `http://www.calle.com/world/PO/`).

For the network domain, we use the following classes:

Internet domains: whois data about DNS domains. For Portugal, the domains database of *Fundação para a Computação Científica Nacional* (FCCN), the managing organization of the ".PT" TLD (Top Level Domain), provides the information used in GKB.

Web Sites: addresses of web sites and their IP addresses. For Portugal, this information is obtained from the Versus web meta-data repository of the tumba! search engine.

2.1. The Concept of Feature and The Information Model

In GKB, we distinguish the name and the feature (or entity) that it represents. We use the notion of feature defined in ISO 19109, "a meaningful object in the selected domain of discourse" [ISO19109 2005]. In the geographic domain, countries, cities and municipalities are examples of such objects. In GKB, features and their names are distinct classes

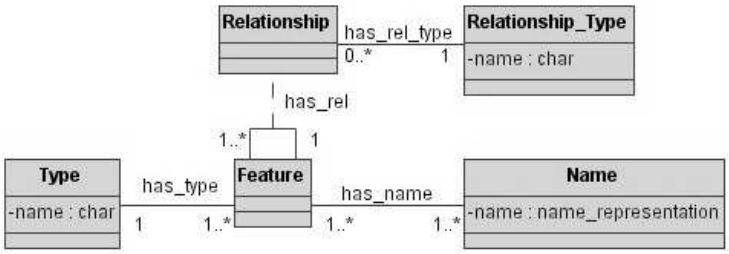

Figure 2. GKB information meta-model

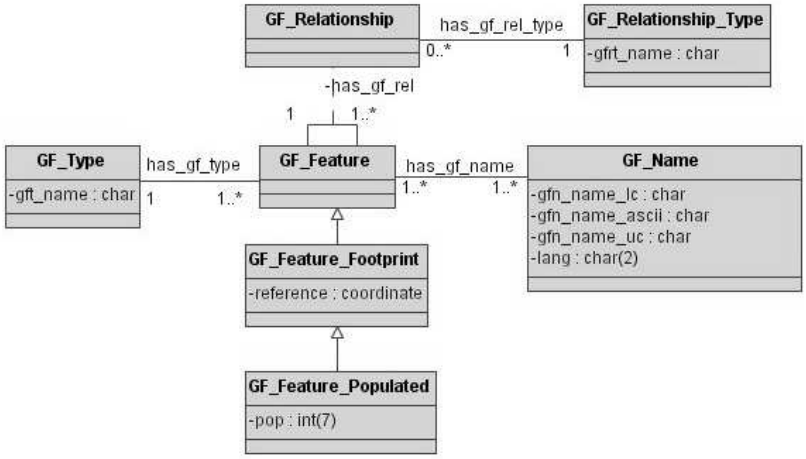

Figure 3. Geographic domain class diagram

and each feature is associated to a feature type. As in ISO 19109, features are classified into feature types on the basis of common sets of characteristics or properties. This approach enables GKB to support many-to-one relationships between names and features. This flexibility also allows the incorporation of new kinds of data. The GKB meta-model is sufficiently generic to represent information from any domain. Figure 2 shows the base information model of GKB. A feature is composed by a name, a type and an information source. A *Feature* has a *Type*, defined in a class, whose instances represent all the feature types identified in information sources. The class *Name* has names identified for every feature in all available information sources. Finally, the classes *Relationship* and *Relationship_Type* capture relationships among features.

The UML class diagram in Figure 2 represents the common meta-model for storing the information held in a repository (or instance) of GKB. This meta-model is then specialized for each information domain supported. The geographic domain is represented in Figure 3. The classes *GF_Type*, *GF_Feature*, *GF_Relationship*, *GF_Name* and *GF_Relationship_Type* represent the same classes of the base meta-model presented in Figure 2. The geographic feature types include municipalities, streets and postal codes. The geographic relationship types are defined as `partOf` and `adjacency`. Geographic features are specialized when we need to capture detailed administrative data, such as population of some regions or geographic coordinates, such as latitude and longitude (*GF_Feature_Populated* and *GF_Feature_Footprint* are examples of class in this category). The class *GF_Name* holds alternative names (names often used with the same

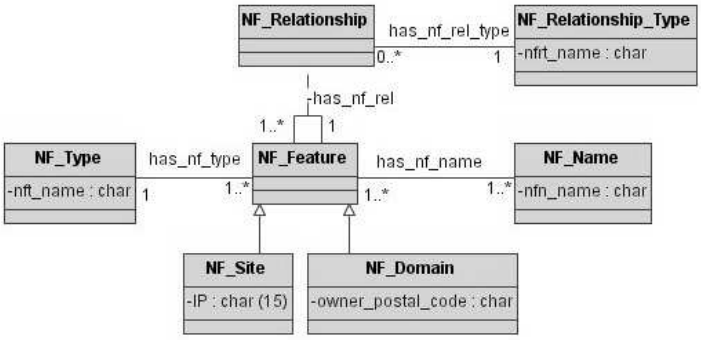

Figure 4. Network domain class diagram

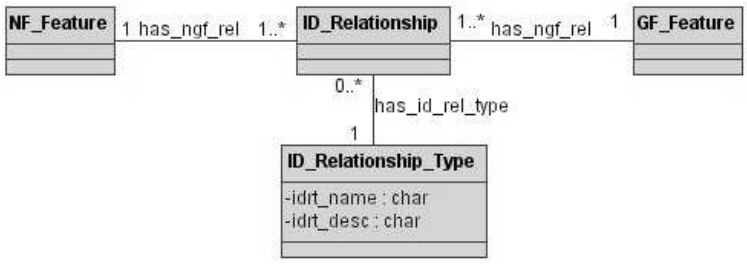

Figure 5. Inter-domain relationships class diagram

meaning of the standard name). For instance, the administrative region of *Nossa Senhora da Conceição* in *Lisboa* is also referenced with the alternative name *Conceição*. This alternative name is associated with the standard name in *GF_Feature*, once it is identified with the same identifier of the standard name. Alternative names have also been considered in another work [Jones et al. 2003]. The *GF_Name* class also stores the language of each name in the attribute `lang`.

Figure 4 represents the network domain class diagram. The class *NF_Type* represents feature types such as *domain* and *site*. The class *NF_Site* specializes the class *NF_Feature* and stores the IP address of the each site, while the class *NF_Domain*, also a specialization of the class *NF_Feature*, stores the web domain owners' postal code. In GKB, we use postal codes to infer inter-domain relationships between geographic and network domains.

Inter-domain relationships between *GF* and *NF* are modeled as shown in Figure 5. One of GKB's applications assigns geographic scopes to web pages. In GKB, a scope is modeled as an inter-domain relationship between a web domain and a geographic feature. For instance, the geographic scope of the web site of the Lisbon municipality, *www.cm-lisboa.pt*, is the city of Lisbon.

3. Data and Knowledge Integration

In general, GKB data sources are independently developed and maintained. They are also designed to serve specific needs. This originates redundancy and a large heterogeneity in terms of information models. Some of them complement each other by providing additional information about a geographical entity. Thus, duplicate information has to be purged out and complementary information should be consolidated to achieve a consistent

view of real world entities. Whenever a new information source is loaded into GKB, we first attempt to detect if the new features are already defined. In that case, we only add new names or geographic relationships to the existing hierarchy.

3.1. Data Cleaning

The process of data cleaning is essential to build a consistent KB. It has three phases, known as ETL (extraction, transformation and loading) [Rahm and Do 2000]. In GKB, we trace the source of each feature and relationship, so we can later assign a level of confidence to each fact. Rahm and Do classify data cleaning problems as single-source and multi-source problems. When cleaning geographic data, we face problems from both classes, which we detail in the rest of this section.

3.1.1. Single-source problems

The most common single-source problems are:

Spelling Errors: Spelling errors are inevitable in large information sources containing data typed by humans. Most of GKB information sources are curated, but errors are still common. The removal of all spelling errors is an impossible task. When detected, such errors are eliminated, but some will always remain.

Invalid Postal Codes: Domain registrars frequently insert invalid postal codes in the network domains database. We detect them when we search a given postal code and cannot validate it. GKB scripts can occasionally detect and correct some of them with the following processing:

- identify sequences of digits in postal code fields in data source being loaded;
- convert the digits to the standard postal format (in Portugal, 4+3 digits, as in "1250-212");
- the digits are considered a valid postal code if the obtained code matches one in the postal codes database.

Insertion of Alternative Names: Data from the gazetteers provide names of localities with and without accented characters as alternatives. When the data source is Portuguese, we just consider the names with accented characters and assume that the others represent alternatives for character encodings that do not support accents. In general, however, it is common to find in a gazetteer alternative place names for places and regions. For example, *São João*, located in *Viana do Castelo*, has the following alternative names: *Vila Chã* and *São João Baptista*. These are stored as alternative names associated to the preferred name with an `equivalent to` relationship.

Correction of Geographic Coordinates: In a gazetteer, a region is sometimes associated with more than one geographic coordinate. In such cases, our default approach is to take the average of the coordinates.

3.1.2. Multi-source problems

We also find inconsistencies when integrating data from multiple information sources. In general, geographic names data sources also organize information in different ways and

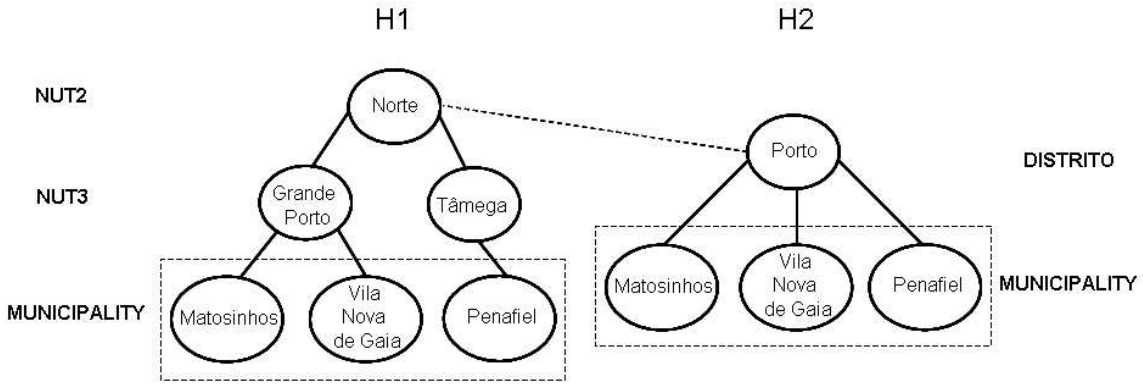

Figure 6. GKB hierarchy from different information sources

we need to address structural heterogeneities as well. To resolve some of these inconsistencies, we assign a level of authority to each GKB information source and use this authority information to resolve inconsistencies in data. For instance, when matching data from CTT and the gazetteer `calle.com`, we found 11 *distritos* in CTT, covering to the 2 Autonomous Regions (Azores and Madeira) in the gazetteer. We assigned to the locations in the gazetteer the corresponding district names in CTT, in fact considering CTT as a more important authority. However, the problem is in general more complex, as the information sources may not be exhaustive or authoritative. In these situations, GKB can keep all the information received for loading, while leaving to its information consumers the possibility of tracing the data origin to their sources and make the final decision about its validity.

3.2. Knowledge Integration in GKB

GKB receives information from multiple sources, each one with knowledge organized differently and representing geographic information at different levels of abstraction. Some sources provide information just about the main regions of a country, while others include feature names down to the level of streets and postal codes. We need to deal with this knowledge in a consistent way. Figure 6 shows a concrete example of a situation where we need to apply our procedure for merging hierarchies in GKB. We have a hierarchy `H1` loaded in GKB and another hierarchy `H2` to be loaded. In `H1`, we have three regions of Portugal: two NUT (*Nomenclatura de Unidade Territorial*) feature types and a narrower type (Municipality). In `H2`, we have two regions of Portugal: `Distrito` and `Municipality` feature types.

Our algorithm merges hierarchies through the following steps (examples given in parenthesis refer to Figure 6): at first, it searches the lowest common features types in both hierarchies (municipality). If it holds, it identifies the common instances between the hierarchies (*Matozinhos*, *Vila Nova de Gaia* and *Penafiel*). Once the common instances are identified, it goes up the hierarchy and searches for the lowest common ancestor (*Norte* in `H1` and *Porto* in `H2`). After these steps, the algorithm verifies the distance (in number of relationships `partOf`) between the common instances of the features types and its ancestors. The ancestor (`Porto`), which has the small distance up to the common instances is merged through a relationship `partOf` with the ancestor (`Norte`) in the another hierarchy. The existing relationships in both hierarchies are maintained. Figure 7 shows the

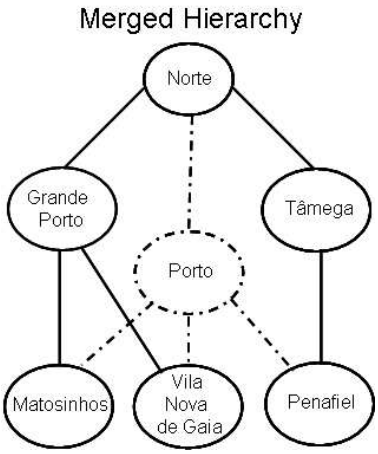

Figure 7. Merged GKB hierarchy

```
geoFeatureName(270,‘‘santiagodocacem’’).
geoFeatureName(270,‘‘santiagocacem’’).
geoFeatureName(270,‘‘santiago-do-cacem’’).
geoFeatureName(270,‘‘santiago-cacem’’).
geoFeatureType(270,‘‘CON’’).
netSiteSubDomain(33684,‘‘www’’).
netSitePrefix(33684,‘‘cm’’).
netSiteDomainToken(33684,‘‘santiago-do-cacem’’).
netSiteTLD(33684,‘‘pt’’).
```

Figure 8. ABox in DLs for the city of “Santiago do Cacém” (the numeric values 270 and 33684 correspond to the feature identifier in an instance of GKB holding these data)

merged hierarchy.

3.3. Using Geographic Knowledge in GKB

GKB not only manages geographic and geographic-related entities and relationships, but also rules relating them. Rules can be added manually or may be automatically inferred by external text mining tools. Rules may also be used by GKB programs to verify domain integrity rules and generate new relationships. To generate relationships, GKB receives the geographic data and rules in order to produce new relationships to be added to the relational database.

In general, the name given to a feature is represented in different ways, depending on the information domain under consideration. For instance, names may be composed of multiple words. In the geographic domains, the space character is the separator; however, in the network domain, this character is invalid in URLs.

Figure 8 shows an extract of the world description of GKB (ABox) in Description Logics. The world description is composed by the different representations of geographic names. Names of the URLs are used in original format, just decomposed by the correspondent domain division. A geographic name encoded in an URL has no spaces or may have hifens substituting for them or still may not have prepositions in its name.

The different representations of the name `Santiago do Cacém` (see the values of the atomic concept `geoFeatureName`) illustrate the ways that we represent the geographic knowledge in DL. The value of the atomic concept `geoFeatureType` corresponds to the geographic type of the name and *270* is the feature's identifier.

For network domain, we represent the URL of sites tokenized in three atomic concepts: subdomnain, domain and top level domain (TLD). In addition, we also create the atomic concept `netSitePrefix`, which indicates the prefix to be used in a rule. For example, `www.cm-santiago-do-cacem.pt` is coded as `netSiteSubDomain(33684,''www'')`, `netSitePrefix(33684,''cm'')`, `netSiteDomainToken(33684,''santiago-do-cacem'')` and `netSiteTLD(33684,''pt'')`, where `33684` is the feature's identifier.

New knowledge is incorporated in GKB through rules, described in the Terminology Description (TBox in DLs): In Portugal, many of the web sites of municipalities are housed in domains whose names contain the prefixes "cm-" or "mun-". We express this knowledge by the following rule:

Municipalities: `hasScope(idN,idG)` $\equiv$ $\exists$`netSiteDomainToken(idN,X)` $\sqcap$
($\exists$`netSitePrefix(idN,''cm'')` $\sqcup$ $\exists$`netSitePrefix(idN,''mun''))` $\sqcap$
$\exists$`geoFeatureType(idG,''CON'')` $\sqcap$ $\exists$`geoFeatureName(idG,X).`

meaning that exits a `netSiteDomainToken X` which has the `netSitePrefixes` "cm" or "mun" and a `geoFeatureType` "CON" with the `geoFeatureName` `X`. When in this rule a matching is found between the values `X` from `netSiteDomainToken` and `geoFeatureName`, we assign that the network feature represented by value `idN` has the geographic scope the feature represented by the identifier `idG`.

Table 1 presents statistics about some of the sites for which we created rules like the above. The number of sites identified for each type and the number of matches ob-

Table 1. Rule-based assigned scopes by GKB to sites of Portugal

Site Type	# of sites	# of matches	Site Type	# of sites	# of matches
distritos	33	17 (52%)	basic schools	1955	124 (6%)
municipalities	288	261 (90%)	training centers	152	55 (36%)
freguesias	300	124 (41%)	high schools	402	105 (26%)

tained after the application of the rules are shown. For instance, Portugal has 308 municipalities and 288 of them have web sites. For these, we get to assign a geographic scope to 261. This simple set of rules can assign geographic scopes to 22% of the site types considered.

We could assign scopes to most of the sites matching the rules above. However, these matchings do not always work because the domain name for some of the sites does is not directly derived from the name of the corresponding feature. For instance, the site `www.cm-ofrades.com` is about the municipality `Oliveira de Frades`.

4. GKB as an Ontology

The information stored in GKB repository can be extracted with a tool named GOG - GKB Ontology Generator. GOG enables selecting parts of the information stored in a

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<rdf:RDF xmlns:gn = "http://xldb.di.fc.ul.pt/geo_net_pt01.owl#">
<gn:Geo_Feature rdf:ID="GEO_238">
  <gn:geo_id>238</gn:geo_id>
  <gn:geo_name xml:lang="pt">Porto</gn:geo_name>
  <gn:geo_type_id rdf:resource="#CON"/>
  <gn:info_source_id rdf:resource="#INE"/>
  <gn:related_to>
      <rdf:Bag>
         <rdf:li>
            <gn:Geo_Relationship>
                <gn:rel_type_id rdf:resource="#PRT"/>
                <gn:geo_id>
                    <rdf:Bag>
                      <rdf:li rdf:resource="#GEO_130"/>
                      <rdf:li rdf:resource="#GEO_3967"/>
                    </rdf:Bag>
                </gn:geo_id>
            </gn:Geo_Relationship>
         </rdf:li>
         <rdf:li>
            <gn:Geo_Relationship>
                <gn:rel_type_id rdf:resource="#ADJ"/>
                <gn:geo_id>
                    <rdf:Bag>
                      <rdf:li rdf:resource="#GEO_127"/>
                      <rdf:li rdf:resource="#GEO_156"/>
                      <rdf:li rdf:resource="#GEO_162"/>
                      <rdf:li rdf:resource="#GEO_331"/>
                    </rdf:Bag>
                </gn:geo_id>
            </gn:Geo_Relationship>
         </rdf:li>
      </rdf:Bag>
  </gn:related_to>
  <gn:population>263131</gn:population>
</gn:Geo_Feature>
</rdf:RDF>
```

Figure 9. An excerpt of GKB-extracted ontology with data about Portugal

GKB instance. The GKB repositories have currently about 0.5 million of features and the user rarely wants to receive full information.

GOG exports the information in the OWL format, a vocabulary extension of RDF (`http://www.w3.org/TR/REC-rdf-syntax/`). Figure 9 presents an excerpt from an instance extracted from GKB. It describes the feature type *Concelho* (abbreviated as `CON`) named *Porto*, which has identifier *GEO_238*. This feature was imported from *Instituto Nacional de Estatística* (`INE`). The *Concelho* of *Porto* has two type relationships with other features: `parteOf` (`PRT`) with features *Grande Porto* and the *Distrito* of *Porto*, identified by codes `GEO_130` and `GEO_3967`, respectively; *adjacency* (`ADJ`) with the features *Gondomar*, *Maia*, *Matosinhos* e *Vila Nova de Gaia*, identified by codes `GEO_127`, `GEO_156`, `GEO_162` and `GEO_331`, respectively. The population of the *Concelho Porto* is `263131` people.

The GKB ontology was validated by RDF Validator (`http://www.w3.org/RDF/Validator/`). The full geographic ontology of Portugal contains more than 418,000 features and we give it as a public resource [Chaves et al. 2005].

In addition to this geographic ontology of Portugal, we generated an ontology of geographic names of the World, obtained by integrating information from public data

```
<gn:Geo_Feature rdf:ID="GEO_170">
  <gn:geo_id>170</gn:geo_id>
  <gn:geo_name xml:lang="en">Guatemala City</gn:geo_name>
  <gn:common_name>
    <rdf:Bag>
      <rdf:li>
        <gn:Geo_Name>
          <gn:geo_name xml:lang="es">Ciudad de Guatemala</gn:geo_name>
        </gn:Geo_Name>
      </rdf:li>
      <rdf:li>
        <gn:Geo_Name>
          <gn:geo_name xml:lang="de">Guatemala-Stadt</gn:geo_name>
        </gn:Geo_Name>
      </rdf:li>
      <rdf:li>
        <gn:Geo_Name>
          <gn:geo_name xml:lang="pt">Cidade da Guatemala</gn:geo_name>
        </gn:Geo_Name>
      </rdf:li>
    </rdf:Bag>
  </gn:common_name>
  <gn:geo_type_id rdf:resource="http://xldb.di.fc.ul.pt/geo-net.owl#CITY-CAP"/>
  <gn:related_to>
     <gn:Geo_Relationship>
         <gn:rel_type_id rdf:resource="http://xldb.di.fc.ul.pt/geo-net.owl#PRT"/>
         <gn:geo_id rdf:resource="#GEO_169"/>
     </gn:Geo_Relationship>
  </gn:related_to>
  <gn:info_source_id rdf:resource="http://xldb.di.fc.ul.pt/geo-net.owl#WIKI"/>
</gn:Geo_Feature>
```

Figure 10. An excerpt of GKB ontology with World data

sources directly available on the Web. Figure 10 presents an excerpt of this ontology, with a description of Guatemala City. This geographic feature is identified by `GEO_170` and its type is `CITY-CAP`. Guatemala City has four common names in English, Portuguese, Spanish, and German. It has a relationship *part-of* (`PRT`) with the feature `GEO_169`, which is declared in another part of the ontology and has the name Guatemala. This information was obtained from `WIKI`, the wikipedia information source.

4.1. Statistics of the Ontologies Created

Table 2 presents descriptive statistics of the ontologies generated from the Portugal and World instances of GKB that we developed. In both ontologies, most of the relationships are of the `partOf` type, while `equivalence` and `adjacency` relationships are much less frequent. The World ontology is much smaller than the ontology about Portugal.

5. Applications using GKB

GKB is one of the components developed under the Geographic Reasoning for Search Engines (GREASE) project (`http://xldb.di.fc.ul.pt/grease`), which researches methods, algorithms and software architectures for assigning geographic scopes to web resources and for retrieving documents using geographical features. Together with GKB, the applications described in this section form the software base used in the GREASE.

GKB is currently used in three different applications which address problems related to classifying and retrieving web pages according to their geographical scope: (1) a

Table 2. Statistics of the Geographic Ontologies of Portugal and World

Statistic	Portugal	World
Number of features	418,065	12,293
Number of relationships	419,867	12,258
Number of part-of relationships	418,340 (99.83%)	12,245 (99,89%)
Number of equivalence relationships	395 (0.09%)	2,501(20,40%)
Number of adjacency relationships	1,132 (0.27%)	13 (0.10%)
Avg. broader features per feature	1.0016	1.07
Avg. narrower features per feature	10.56	475.44
Avg. equivalent features per feature with equivalent	1.99	3.82
Avg. adjacent features per feature with adjacent	3.54	6.5
Number of features without ancestors	3 (0.00%)	1(0.00%)
Number of features without descendants	374,349 (89.54%)	12,045 (97,98%)
Number of features without equivalent	417,867 (99.95%)	11,819 (96,14%)
Number of features without adjacent	417,739 (99.92%)	12,291 (99,99%)

geographical named entity recognition, classification and grounding tool, (2) a document classifier for geographical scopes, and (3) an information retrieval interface for geographical queries.

In language processing, the task of extracting and distinguishing different types of entities in text is usually referred to as Named Entity Recognition (NER) [Kalfoglou and Schorlemmer 2003]. Typical NER systems consist of at least a tokenizer, NE datasets (gazetteers) and NE extraction rules. The rules for NE recognition are the core of the system, combining the named entities in the gazetteer with elements such as capitalization and the surrounding text. Mikheev et al. showed a NER system could perform well even without gazetteers for most classes, although this was not the case for geographical entities [Mikheev et al. 1999]. The same study also showed that simple matching of the input texts to previously generated lists performs reasonably well in this last case, again confirming the need of a good source of geographical place names in order to accurately extract geographical references from textual documents. Our system (tools (1) and (2)) for geographical names uses the information at GKB as the main dataset, together with simple hand-coded rules [Silva et al. 2004]. It associates the found entities to the corresponding GKB feature, so that subsequent processing operations can reuse the GKB ontology to infer extra knowledge.

Assigning geographical scopes to documents is a very difficult classification problem, leaving open challenges to current machine learning (ML) approaches. For instance, the number of occurrences of a given geographical name is insufficient to base probabilistic methods on, leading to the failure of typical methods. Recognizing geographical named entities in a document is also in itself not enough for classification, as geographical entities are ambiguous [Page et al. 1999]. We developed a specific method for this task that instead of the standard ML methodology of automatically inferring classifiers from a training set of documents uses the recognized geographical named entities together with a combination/disambiguation algorithm that builds on the GKB ontological relationships [Martins and Silva]. The disambiguation algorithm sees the ontology as a graph and takes its inspiration on PageRank [Baeza-Yates and Davis 2004]. The ge-

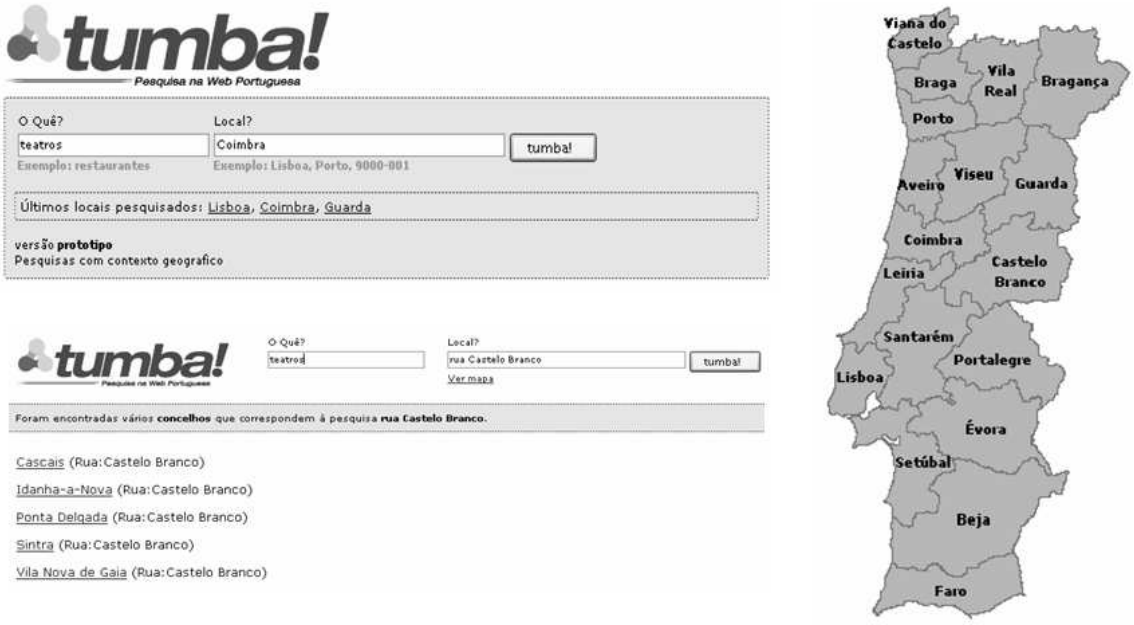

Figure 11. Examples of the interfaces for geographic IR using GKB

ographical features and the ontological relationships between them can be seen as the nodes/vertexes of a graph, and the document occurrence frequency associated with each feature can be used as "relevance" weights. A slightly modified version of the PageRank ranking algorithm is applied to this graph, in order to compute a score for each GKB feature. The highest scoring feature is in the end selected as the geographical scope for the document.

Finally, GKB is also used in the interface of a geographical information retrieval system, assisting users in the formulation of queries. Since geographical names are ambiguous, GKB is used to present users with different alternatives to their queries. Figure 11 presents the Geo-Tumba interface, which was designed to support queries with a defined geographic scope. In the field `Local?` the user types the region, street, postal code or another geographic feature to reduce the scope of the query. When an ambiguous geographic name is detected in the query, Geo-Tumba shows possible alternatives for the user disambiguates its query. For example, the name "rua Castelo Branco" occurs in 5 different municipalities, which are presented in the left inferior side of the Figure 11. Besides the text query, the user can use maps to define the scope of a query.

We are now preparing the search engine tumba! (`http://www.tumba.pt`) to participate on Geo-CLEF (`http://ir.shef.ac.uk/geoclef2005/`). Two of the main challenges of this evaluation are translating locations and finding (or creating) suitable multilingual gazetteer lists. GKB is being used to provide support to the translations of the geographic queries. Presently, it is loaded with data in the Portuguese, English, Spanish and German languages, whose are the languages used in Geo-CLEF.

Our experiences with the three applications described above confirm the advantages and usefulness of using GKB to integrate and share geographical information from different sources.

6. Final Remarks

We presented a domain-independent model for storing geographic knowledge, supporting multiple applications related to geographic IR. The major concerns in the design of GKB were the incorporation of data from distinct information sources and the sharing of the collected knowledge as formal ontologies. We added new knowledge to GKB based on rules, which allow to perform inferences. Finally, we showed extracts and descriptive statistics of the ontologies generated from Portugal and World instances of GKB.

We are in the process of augmenting the knowledge present in this repository with the semantic relationships between geographic entities extracted from the texts of the Portuguese Web. This process should be iterative and progressively expand the knowledge stored in GKB.

Acknowledgments

Marcirio Silveira Chaves is supported by FCT, *Fundação para a Ciência e Tecnologia*, through grant POSI/PLP/43931/2001, co-financed by POSI. Bruno Martins is supported by FCT through grant SFRH-BD-10757-2002. GREASE is a project sponsored by FCT, number POSI/SRI/47071/2002. We thank Daniel Gomes for providing us the meta-data of the web sites of the Portuguese web.

References

Alani, H., Kim, S., Millard, D. E., Weal, M. J., Hall, W., Lewis, P. H., and Shadbolt, N. R. (2003). Automatic Ontology-based Knowledge Extraction from Web Documents. *Intelligent Systems, IEEE*, 18(1):14–21.

Alexa (2005). Alexa. http://pages.alexa.com/prod_serv/data_services.html.

Baader, F., Calvanese, D., Nardi, D., McGuinness, D., and Patel-Schneider, P., editors (2003). *The Description Logic Handbook: Theory,Implementation and Applications*. Cambridge University Press.

Baeza-Yates, R. and Davis, E. (2004). Web page ranking using link attributes. In *Proceedings of WWW-04, the 13th international World Wide Web conference - Alternate track papers & posters*, pages 328–329. ACM Press.

Chaves, M., Martins, B., and Silva, M. J. (2005). Geographic Knowledge Base. DI/FCUL TR 05–12, Department of Informatics, University of Lisbon.

Etzioni, O., Cafarella, M., Downey, D., Popescu, A., Shaked, T., Soderland, S., Weld, D. S., and Yates, A. (2004). Methods for Domain-Independent Information Extraction from the Web: An Experimental Comparison. In *Proc. of the Nineteenth National Conference on Artificial Intelligence - AAAI'04, San Jose, California*, pages 391–398.

Gruhl, D., Chavet, L., Gibson, D., Meyer, J., Pattanayak, P., Tomkins, A., and Zien, J. (2004). How to build a WebFountain: An architecture for very large-scale text analytics. *IBM Systems Journal - Utility Computing*, 43(1).

Irie, R. and Sundheim, B. (2004). Resources for Place Name Analysis. In *Fourth International Conference on Language Resources and Evaluation - LREC2004*, pages 317–320.

ISO19109 (2005). ISO 19109. https://www.seegrid.csiro.au/twiki/pub/Xmml/FeatureModel/19109_DIS2002.pdf.

Jones, C. B., Abdelmoty, A. I., and Fu, G. (2003). Maintaining Ontologies for Geographical Information Retrieval on the Web. In *Proc. of OTM Confederated International Conferences CoopIS, DOA, and OOBASE*, pages 934–951.

Kalfoglou, Y. and Schorlemmer, M. (2003). Ontology mapping: the state of the art. *Knowledge Engineer Review*, 18(1):1–31.

Manov, D., Kiryakov, A., Popov, B., Ognyanoff, D., Kirilov, A., and Goranov, M. (2003). Experiments with Geographic Knowledge for Information Extraction. In *Proc. Workshop on Analysis of Geographic References - Edmonton, Canada*.

Martins, B. and Silva, M. J. A graph-based ranking algorithm for geo-referencing documents. (To Appear).

McGuinness, D. L. and van Harmelen, F. (2004). OWL Web Ontology Language. http://www.w3.org/TR/2004/REC-owl-features-20040210/.

Mikheev, A., Moens, M., and Grover, C. (1999). Named Entity Recognition without Gazetteers. In *Proc. of the Ninth International Conference of the European Chapter of the Association for Computational Linguistics (EACL'99)*, pages 1–8, Bergen, Norway.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The PageRank Citation Ranking: Bringing Order to the Web. Technical Report SIDL-WP-1999-0120, Stanford Digital Library.

Rahm, E. and Do, H. H. (2000). IEEE Bulletin of the Technical Committee on Data Engineering. *Data Cleaning: Problems and Current Approaches*, 23(4).

Sheth, A., Aleman-Meza, B., Arpinar, I. B., Bertram, C., Warke, Y., Ramakrishnan, C., Halaschek, C., Anyanwu, K., Avant, D., Arpinar, F. S., and Kochut, K. (2004). Semantic Association Identification and Knowledge Discovery for National Security Applications. *Special issue of Journal of Database Management*.

Silva, M. J., Martins, B., Chaves, M., Cardoso, N., and Afonso, A. P. (2004). Adding Geographic Scopes to Web Resources. In *ACM SIGIR 2004 Workshop on Geographic Information Retrieval, Sheffield, UK*.

Tudhope, D., Alani, H., and Jones, C. (2001). Augmenting Thesaurus Relationships: Possibilities for Retrieval. *International Journal on Computer Science and Information Systems*, 1(8).