GKB - Geographic Knowledge Base

Marcirio Silveira Chaves, Mário J. Silva
and Bruno Martins

DI–FCUL TR–05–12

July 2005

Departamento de Informática
Faculdade de Ciências da Universidade de Lisboa
Campo Grande, 1749–016 Lisboa
Portugal

Technical reports are available at `http://www.di.fc.ul.pt/tech-reports`. The files are stored in PDF, with the report number as filename. Alternatively, reports are available by post from the above address.

Contents

List of Tables

List of Figures

GKB - Geographic Knowledge Base

Marcirio Silveira Chaves, Mário J. Silva and Bruno Martins
Department of Informatics
Faculty of Sciences - University of Lisbon
1749-016 Lisbon, Portugal

July 2005

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 the repository. GKB includes tools for generating ontologies, which are being used by multiple Semantic Web applications. In addition GKB supports multiple languages. To illustrate how it is being used, we describe some applications that interact with the repository or load the generated ontologies.

1 Introduction

The vision of the Semantic Web is a distributed system for knowledge representation and computing. However, a barrier to its success is the need for annotated resources in a standardised machine understandable format. For instance, the natural language sentence "Lisbon is the capital of Portugal" should be annotated with a formal representation of it, e.g. Lisbon can be annotated as "city", Portugal as "country", and the sentence should be annotated through a structure "related-to(sentence, Lisbon); part-of(Lisbon, Portugal)".

In this context, a knowledge base (KB) has a key role in supporting applications such as Text Mining, Information Retrieval and Natural Language Processing. An interesting text mining problem concerns the fact that many information resources on the Web are primarily relevant to geographically limited communities. Finding automatic ways of assigning geographical scopes to these resources ("geo-referencing" Web documents) is a challenging problem, getting increasing attention from text mining researchers [Manov et al., 2003, Purves and Jones, 2004]. A geographic knowledge base provides support for query disambiguation, query term expansion, relevance ranking, document annotation, and reasoning about geographical concepts.

In this technical report, we present the Geographic Knowledge Base (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. In GREASE, the main purpose of GKB is to provide a common place for integrating data from multiple sources (not necessarily disjoint and each with their own peculiar formats) under a common schema, supporting mechanisms for storing, maintaining, and exporting the assembled knowledge about geographic entities and Web resources. These

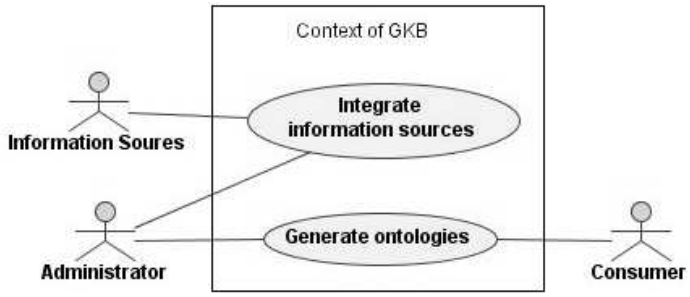

Figure 1: Context of GKB

data are then used by several applications that require geographic knowledge. We represent the geographic and network knowledge in Description Logics (DL) [Baader et al., 2003], which allows us expanding the knowledge ever stored. DL is the formalism under the languages used in Semantic Web.

Most geographic information stored electronically is still embedded in legacy databases and file storage formats that were not designed to interact easily with other software. One of the contributions of this work is provide information in a machine readable format in the context of Semantic Web [Berners-Lee et al., 2001]. So, the development of tools to transfer this information for a Web suitable format is a requirement. We are concerned about two of the main challenges dealing with ontologies: their creation and management. Design, development, storage and maintenance are discussed in this paper.

GKB is implemented on top of a relational database, which maintains the consolidated information collected from various information sources. In addition to the database, GKB has two sets of tools:

- converters, which load data from the various source formats into GKB. These tools perform some amount of data normalisation in order to maintain a single unified view of all the information. Converters to this same task are also used in [Hill, 2000];
- generators, which output the GKB contents as ontologies. The generated ontologies are represented in the OWL (Web Ontology Language) standard [McGuinness and van Harmelen, 2004], suitable to be used by other components of the GREASE project.

The ontologies generated are mainly used for Information Retrieval and Information Extraction. Other components developed in the GREASE project such as CAGE (acronym to CApturing Geographic Entities - a geographic entity names recogniser) and Geo-Tumba (a geographic search engine) use the information in GKB. These modules will be latter introduced in Section 9. We intend to augment the knowledge present in this repository by exploring the semantic relations among the geographic entities described in the texts of the Portuguese Web. The search by complex semantic relationships between semantically annotated entities is the next step of the Semantic Web [Sheth et al., 2004].

1.1 Context and Requirements of GKB

Figure 1 presents the context of GKB. GKB has two main use cases, *Integrate Information*

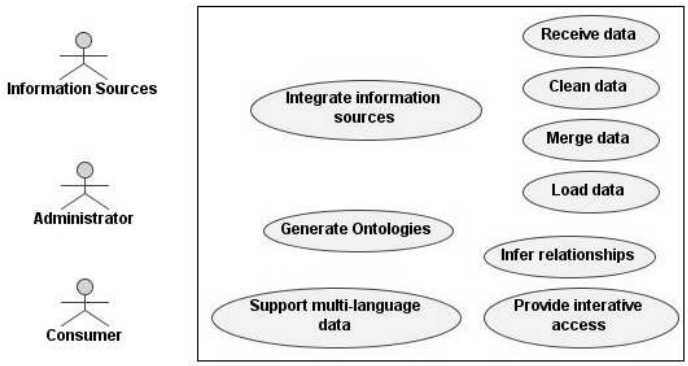

Figure 2: Requirements of GKB

Sources and *Generate Ontologies*. An *Administrator* is responsible for managing GKB. Ontologies are produced for *Consumers*, which include tools and software systems such as CAGE and Geo-Tumba.

Figure 2 shows the requirements of GKB. In detail, these are:

Integrate information sources: GKB must support the addition of data from many information sources and different suppliers quickly. Some of the sources update information frequently. Public administration reorganisation or the creation of new urban areas are examples of events that trigger updates to information present in GKB.

Receive data: The administrator receives data in text format and studies its characteristics. Based on this preliminary analysis, he implements cleaning scripts to import the information source data into the GKB.

Clean data: A rigorous process before the insertion of data into GKB is the cleaning of the imported data, which usually has distinct formats and can be replicated. After the administrator receives the data, he modifies the existing scripts or creates new ones to clean these data. This process includes the resolution of inconsistencies in received data.

Merge data: Some of the information sources have data in common. The administrator should identify these cases and implement scripts to merge the data before the loading phase. This use case help us to avoid inserting duplicate data in GKB.

Load data: The data from an information source, which may have previously been merged with other sources is added to the data in the GKB.

Infer relationships: Inferences can be performed when the same data is associated with different entities. For example, postal codes are present both in geographic and network domains. The administrator runs programs that add relationships between GKB entities derived from the implicit knowledge about these information domains.

Generate ontologies: The administrator creates the ontologies to the consumers.

Provide interactive access: GKB should provide interactive access to the data. Users and the administrator use a Web interface to access interactively the information.

Support multi-language data: GKB stores multi-language data. It should support this kind of characteristic to allow users and applications to choose the assignation of geographic names in a specific language or all of them.

The remainder of this paper is organised as follows: the next Section discusses related works. Section 3 presents the information domains and the conceptual design of GKB. GKB is instanced with Portugal and World data, which are described in Sections 4 and 5, respectively. Section 6 discusses the data quality and information integration issues. Section 7 presents rules in DL used to expand the knowledge in GKB. Section 8 describes GKB as an ontology. Section 9 introduces the applications using GKB. Finally, Section 10 presents the final conclusions and some ideas to future work.

2 Related Work

We split this section in structures to represent knowledge and geographic knowledge representation. In the former, we introduce the main concepts used in our work. In the latter, we present some of the works that deal with geographic knowledge representation.

2.1 Structures to Represent Knowledge

Frequently, concepts like gazetteer, ontology and thesaurus are used indistinctly. We present our understanding about them in the following.

A simple definition to gazetteer can be found in WordNet (`http://www.cogsci.princeton.edu/`): a geographical dictionary. Wikipedia (`http://en.wikipedia.org/wiki/Gazetteer`) extends this definition asserting that a gazetteer typically contains information concerning the geographic makeup of a country or region, and the social statistics. We add that often a gazetteer includes information about latitude and longitude coordinates.

The most used definition of an ontology is given by Gruber: an ontology is an explicit specification of a conceptualisation [Gruber, 1993]. Fensel details this definition, asserting that a conceptualisation refers to an abstract model of some phenomenon in the world, which identify relevant concepts from that phenomenon [Fensel, 2001]. A conceptualisation explains the intended meaning of the terms used to indicate relevant relations [Guarino, 1997]. The restriction to be explicit means that the concepts and relations between them are explicitly defined, that is, there is no ambiguity in the ontology.

Gonzalez discusses more than 15 definitions of a thesaurus [Gonzalez, 2001]. According to him, a thesaurus is a lexical database that implements an ontology, where the described terms should be interpreted as concepts and its relationships constitute the essence of the description of these concepts. A thesaurus is a closed language restricted normally by three relationships, which are: equivalence (Used For and USE), hierarchy (Broader Term/Narrower Term) and associativity (Related Term) [ISO2788, 1986].

According to the definitions of a gazetteer, it is a flat structure, which does not contain explicit relationships between the terms. On the contrary, with an ontology, we capture and formalise the main concepts and their relationships in a knowledge domain (in our case, the geographic one). We can use an open language to build an ontology, without restrictions about kinds of relationships as in a thesaurus.

Other concepts related to these presented here, such as taxonomies and controlled vocabularies are available from `http://www.ontopia.net/topicmaps/materials/tm-vs-thesauri.html`.

2.2 Geographic Knowledge Representation

Geographic knowledge bases have been used lately to support research in Information Retrieval and Geographical Information Management [Inoue et al., 2002, Jones et al., 2003, Fu et al., 2003, Gravano et al., 2003, Markowetz et al., 2004].

Manov et al. worked in the creation of a structured knowledge base according to an ontology, instead of having 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 our approach, instead of obtaining data only from gazetteers, we also import data from other classes of information sources. We followed a similar approach to Alani et al., relying 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 use of ontology(ies) to represent the content expressed in documents or databases has been presented in several works [Mena, 1998, Kietz et al., 2000, Nobécourt, 2000, Cruz et al., 2002, Szulman et al., 2002, Hyvönen et al., 2004, Purves and Jones, 2004].

The main aims identified to use ontology(ies) to represent knowledge are:

- to aid tasks such as search disambiguation, expansion of terms in the search, sort of relevance and annotation of the Web resources [Fu et al., 2003].
- to create a structured KB instead of having this base according to somewhat flat structures of gazetteer lists [Manov et al., 2003].

Spatially-Aware Information Retrieval on the Internet (SPIRIT) is a European project which aim is use ontologies to aid task such as query disambiguation, query term expansion, ranking and annotation of Web resources [Fu et al., 2003]. The main characteristic of this system is the production of ways to facilitate and support queries using terms and geographic relations.

An immediate question that arises when we mention geographic systems is the use of gazetteers. Fu, Abdelmoty and Jones consider problems with their use such as the lack of sufficient support to explicitly code spatial relationships (overlap and adjacency, for example) and use of geographic attributes of different kinds coded in the same way [Fu et al., 2003]. In order to deal with these and other problems, SPIRIT uses a geographic ontology represented in the DAML+OIL language.

Gravano, Hatzivassiloglou and Lichtenstein classify queries according to the geographical localities [Gravano et al., 2003]. Specifically, a query can belong to one of the two categories: global or local. A query is considered local when it is composed by a geographic term, otherwise it is global. A database of 1,605 names of the main locations in USA was used in order to help the identification of the geographic scope of the query. A deficiency in their classification is the lack of a phase of disambiguation of terms. A document where the term *Washington* occurs is classified as located in the city of *Washington*, independently of this term referring to a name of a person.

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

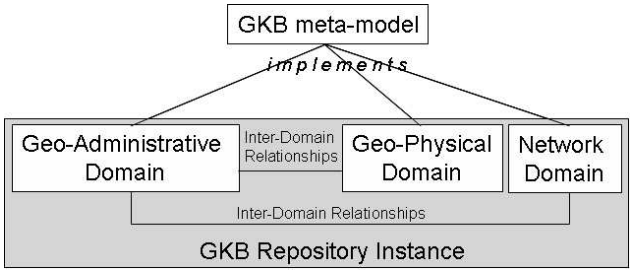

Figure 3: GKB information architecture

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.

3 Conceptual Design of GKB

Figure 3 shows the information architecture of GKB. Data is organised 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.

Ontological relationships among the features of each domain are also described in both repositories. 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 geographic features, specialising 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 site as a relationship between the Web site (a network domain entity) and a geographic region (a geographic domain entity).

3.1 Features and the Meta-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 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.

The meta-model presented in Figure 4 shows the common meta-model for storing the information held in 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

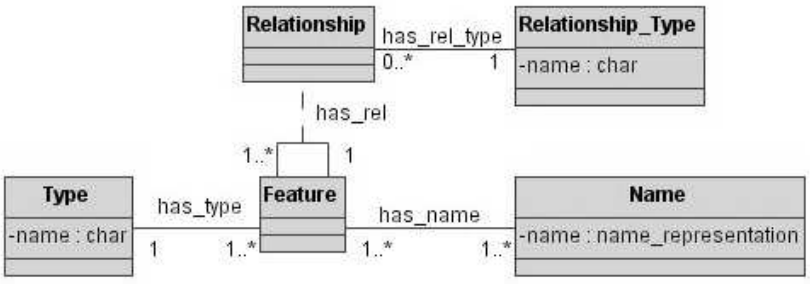

Figure 4: GKB information meta-model

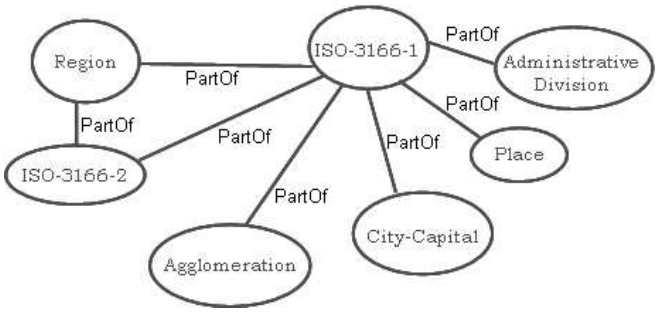

Figure 5: Feature types dependencies for the geo-administrative domain of GKB (instance of the World)

all available information sources. Finally, the classes *Relationship* and *Relationship_Type* capture relationships among features.

We have developed two GKB instances: one stores data about the geo-administrative and network domains of Portugal and another with geo-administrative and geo-physical data about the World. For Portugal, we have more detailed information, including the main administrative regions (names of streets, for example) and their geographic coordinates. The network domain has information about DNS (Domain Name Service) domains hosting Web sites of the Portuguese Web. The second repository stores data about the geographic domain, which is split in administrative and physical.

3.2 Feature Types Dependencies

For each GKB instance, we define a set of feature types and relationship types. In the integration process, we use the implicit knowledge about the dependencies between the feature types defined for each instance. Figure 5 presents these dependencies for the geo-administrative domain, which was the first loaded in GKB with World data.

The main feature type is `ISO-3166-1`, which encompasses the countries and territories around the world. The feature types Region, `ISO-3166-2`, `Agglomeration`, `City-Capital`, `Place` and `Administrative Division` have a `part of` relationship with ISO-3166-1. It is important to note that the relationship between `ISO-3166-1` and `Region` is bidirectional, that is, a feature of type `ISO-3166-1` can be part of a `Region` (*Nicaragua* is part of *Latin America*) or a `Region` can be part of a `ISO-3166-1` feature (*Siberia* is part of *Russia*).

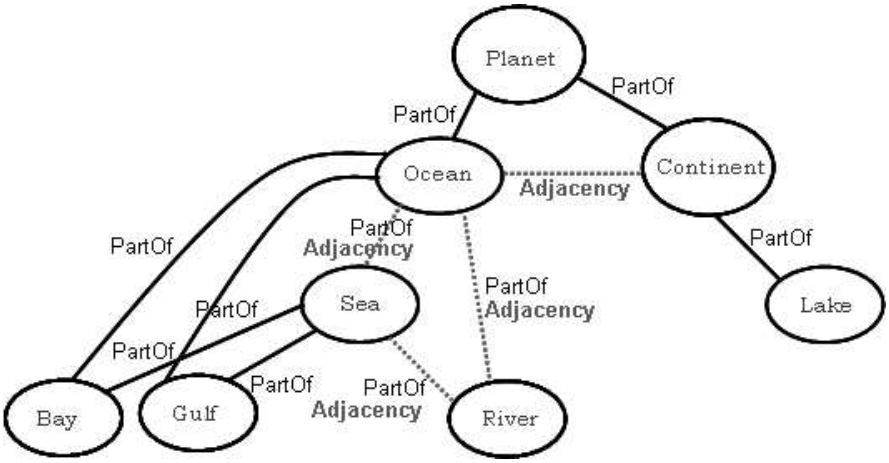

Figure 6: Feature types dependencies for the geo-physical domain of GKB (instance of the World)

We model the dependencies between feature types in the geo-physical domain as represented in Figure 6. The upper feature type is `Planet`, which is composed by `Oceans` and `Continents`. These have a adjacency relationship between them. `Ocean` is related to the feature types `Sea` and `River`. Both them are part of or adjacent to `Ocean`. `River` can be also part of or adjacent to `Sea`. `Bay` and `Gulf` can be part of both, `Sea` or `Ocean`. Finally, `Lake` is part of `Continent`.

In addition to the dependencies above, there are also dependencies between feature types of distinct domains. Figure 7 shows the dependencies between the geo-administrative and geo-physical domains.

The geo-administrative domain is related to the geo-physical domain through the feature types `ISO-3166-1` and `Region`. An instance of an `ISO-3166-1` can be part of `River`, `Continent` or `Lake`. It may be adjacent to `Bay`, `Gulf`, `Sea`, `River` or `Lake`. An instance of a `Region` can be part of a `Planet`, in our case *Earth*.

4 Instance of Portugal

The model of the geographic domain for GKB instance of Portugal is represented in Figure 8. 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 4. The geographic feature types include municipalities, streets and postal codes. The geographic relationship types are defined as `part of` and `adjacency`. Geographic features are specialised when we need to capture detailed administrative data, such as population of some regions or geographic coordinates, such as latitude and longitude. The classes `GF_Feature_Populated` and `GF_Feature_Footprint` are specialisations of the class `GF_Feature`. The `GF_Name` class stores alternative names (names often used with the same meaning of the standard name). For instance, the geo-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 other works [Jones et al., 2003, Hill, 2000].

Figure 9 represents the network domain data model. The class `NF_Type` stores feature types such as *domain* and *site*. The class `NF_Site` specialises the class `NF_Feature` and stores the IP address of the each site, while the class `NF_Domain`, also a specialisation of the class `NF_Feature`,

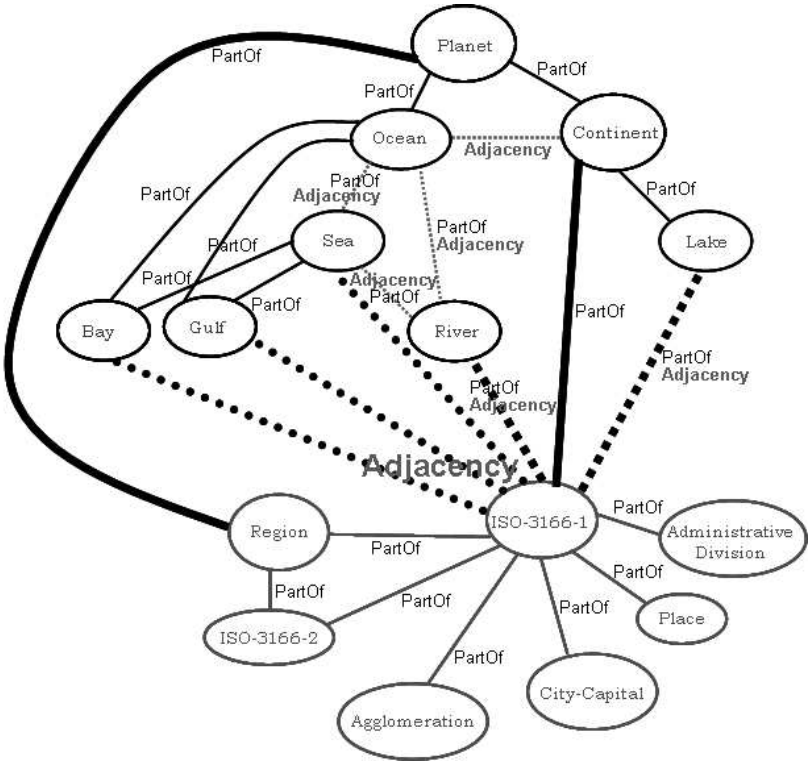

Figure 7: Feature types between the geographic domains

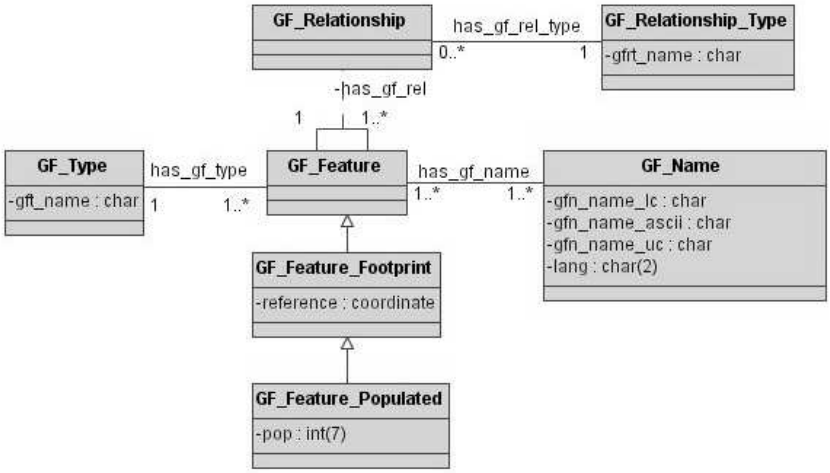

Figure 8: Geographic domain data model

stores the Web domain owners' postal code. We use postal codes to associate features between geographic and network domains.

With the meta-model of Figure 4 we can represent any knowledge domain in GKB. Presently,

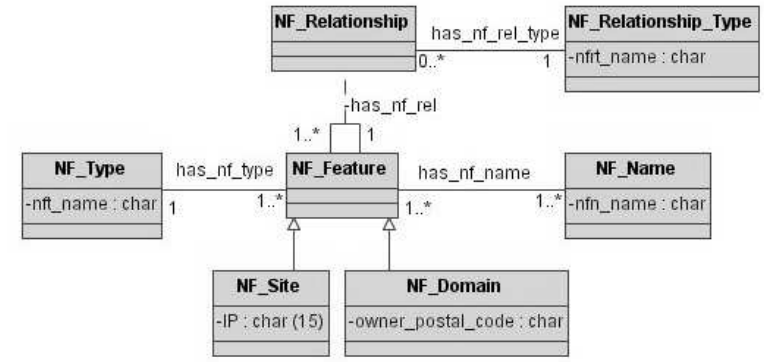

Figure 9: Network domain data model

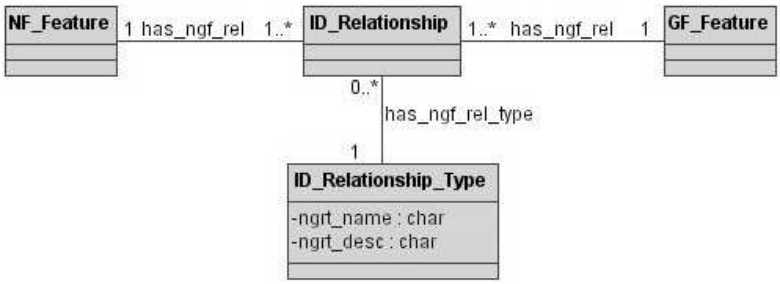

Figure 10: Inter-domain relationships data model

GKB has just two domains, which are inter-related through the data in common. However, we intend to expand GKB to allow the definition of relationships between features from different domains. So, we need to provide a generic and extensible model to support inter-domain relationships, which is presented in Figure 10. The class `ID_Relationship` stores the relationships between the features of the inter-related domains. In GKB, we define a relationship between features in the network and geographic domains.

In the GREASE project, our goal is to assign 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.

Figure 11 presents the full data model for the GKB instance of Portugal. The only class still not previously explained is `Info_Source`, which stores both the name of the sources and the date that it was inserted into GKB. `Info_Source` allows us to version data loaded into GKB over time. Each feature and relationship inserted into GKB is associated to one single information source.

Appendix B presents a SQL script to create a relational schema implementing this model in a MySQL database.

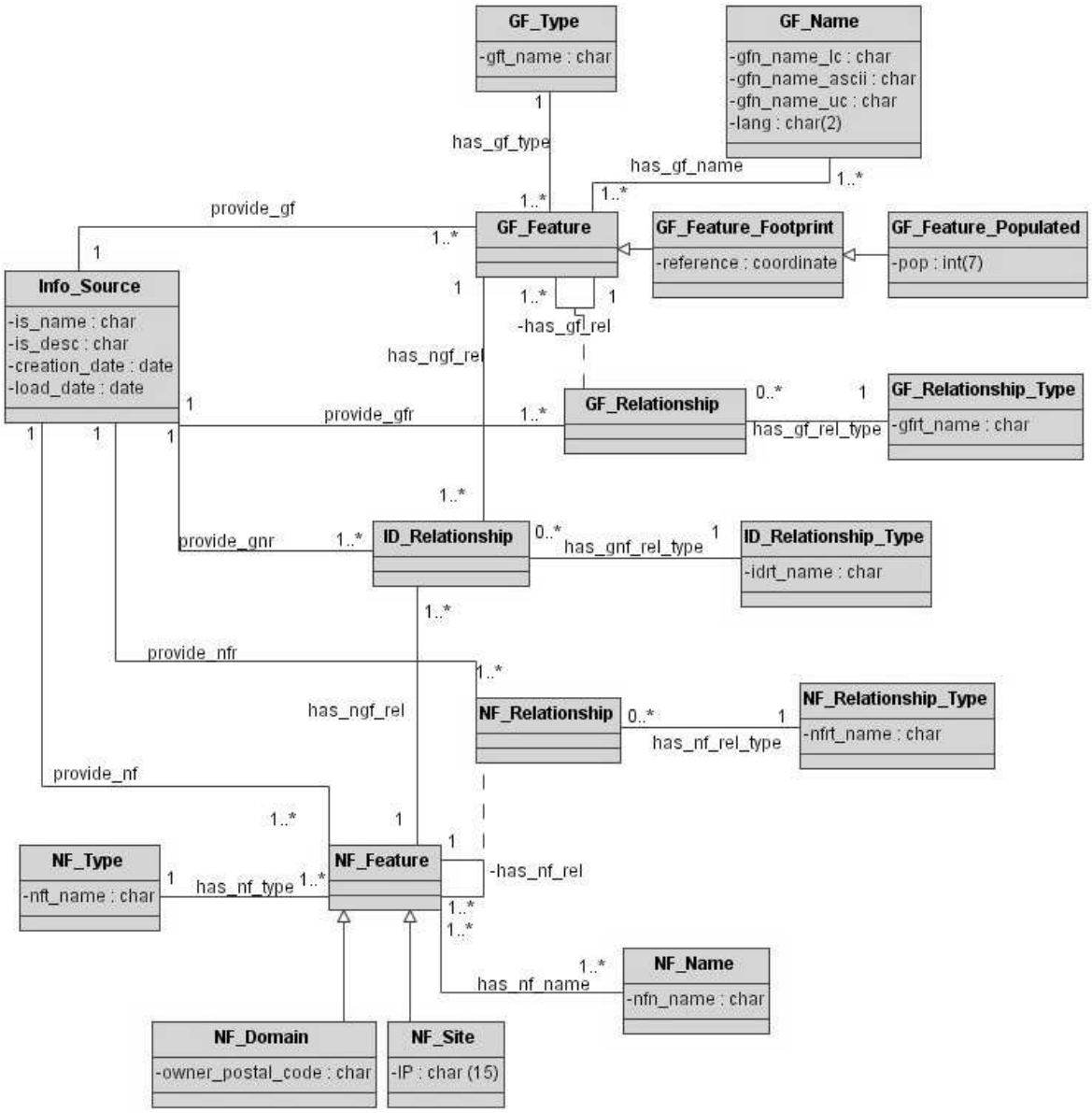

Figure 11: Full data model of GKB (instance of Portugal)

4.1 Information Sources

GKB collects data from several classes of information sources. We next describe each one of the information sources used to load the GKB instance of Portugal in detail.

4.1.1 Geo-administrative Domain

Detailed information must be collected from several classes of information sources. Examples of classes include administrative (statistical), postal, administrative, religious and judicial. This section characterises the information sources presently providing geo-administrative information to GKB.

Administrative (ADM:INE): Databases from the Instituto Nacional de Estatística (INE) concerning demographics and administrative information, such as the territorial division. The INE's administrative classification is mandatary being split by Nomenclature of Territorial Units (NUT). NUT1 (all national territory), NUT2 (*Norte, Centro, Lisboa e Vale do Tejo, Alentejo* and *Algarve* more *Região Autónoma da Madeira* and *Região Autónoma dos Açores*), NUT3 (subregions of NUT2), NUT4 (subregions of NUT3). In addition ANMP (*Associação Nacional de Municípios Portugueses*) provides us the adjacency relationships among the *distritos* and `municipalities`.

Postal (POS:CTT): The Portuguese Post Office (CTT) publishes a database of postal codes. From this database we get, for each postal code, the following types of administrative information: *distrito*, `municipality`, *localidade* and *arruamento*. For instance, the 2775-096 postal code identifies the *arruamento Avenida Infante Dom Henrique* in the *localidade* of *Murtal* in the `municipality` of *Cascais* in the *distrito* of *Lisboa*. The type *arruamento* was subdivided in more specific types according to the occurrences found in POS:CTT. A full list of types present in this domain can be found in Appendix A.

Gazetteer (GEO:GAZ): Directory of cities, towns, and regions in Portugal (http://www.calle.com/world/PO/). The gazetteer provides the geographic coordinates of the main regions of Portugal. An example of instance of such gazetteer is the *distrito* of *Porto*, region of *Matosinhos* with latitude *41.1833° N* and longitude *8.7000° W*. In World level, this gazetteer provided us information about the names of the countries, territories and the main subregions into these countries. We follow the standard ISO-3166-1 and ISO-3166-2 (`http://www.iso.org/iso/en/prods-services/iso3166ma/02iso-3166-code-lists/list-en1.html`). The feature *Bremen*, which is a subregion (*ISO-3166-2*) into *Germany*, is an example of data from GEO:GAZ.

Wikipedia (ADM:WIKI): This is an on-line encyclopedia from which we get the list of *freguesias* and `municipalities`. For instance from ADM:WIKI we obtain the information that the *freguesia* of *Santa Isabel* belongs to the `municipality` of *Lisboa*.

4.1.2 Network Domain

The network domain is composed by data about Web domains and Web sites which come from two information sources:

Web domains (NET:FCCN): *Fundação para a Computação Científica Nacional* (FCCN) domains database. NET:FCCN provides the domains registered under PT top level domain as well as the postal code of the registrant. An instance of NET:FCCN is *igrejacampogrande.pt*, which has the *2670-459* postal code.

Versus (NET:VERSUS): This is a repository of Web metadata, which provided us with the last two crawls of Web sites (PT4 and PT5) performed in the scope of search engine tumba! (`www.tumba.pt`). Each site got from NET:VERSUS is associated to an IP number [Gomes et al., 2002]. The site *www.fc.ul.pt* and its IP *194.117.4.40* is an example of data from this information source.

In addition to these sources, we intend to receive information from Web users related to the new domains, scope of the new or existent domains and so on. The idea is allow the users to contribute with his knowledge, in order to improve the knowledge stored in GKB.

4.2 Loading Procedure

In this section we describe the sequence of steps taken to load the GKB instance of Portugal with data from the information sources described above.

4.2.1 Geo-administrative Domain

Loading of the geographic domain with data about Portugal involves the following steps:

1. Population of the `Info_Source`, `GF_Type` and `GF_Relationship_Type` classes.
2. Population of the `GF_Name` and `GF_Feature` classes with administrative data, namely `NUT1`, `NUT2`, `NUT3` and `municipalities`.
3. Population of the `GF_Relationship` class with data from `GF_Feature` class. All geographic domain relationships are stored in `GF_Relationship`. It is composed by two feature identifiers, a type identifier of the semantic relation and the information source identifier. For example, we search the names into cleaned POS:CTT to store the relation that a `municipality` is part of a *distrito*. We search the feature identifiers of these names in `GF_Feature` class and store the ones found together the type identifier of semantic relation extracted from `GF_Relationship_Type` class into `GF_Relationship` class.
4. Population of the `GF_Footprint` class with data from GEO:GAZ. These data are latitudes and longitudes inserted into `GF_Footprint` class. Each pair of coordinates is associated to a *localidade* in `GF_Feature` class.
5. Population of the `GF_Name` class with alternative names from GEO:GAZ. To load these alternative names into GKB, we first verify if each of these names is not present in the `GF_Name` class. If it is not found, we insert it in both `GF_Name` and `GF_Feature` classes. Subsequently, we get the feature identifiers of both the preferred name and the alternative name and store them in the `GF_Relationship` class with the relation identifier `equivalent` extracted from `GF_Relationship_Type` class.

4.2.2 Network Domain

1. Population of the `NF_Name`, `NF_Feature` and `NF_Site` classes with data from Web sites crawled in PT4 and PT5. The loading of this classes is described in the following:
   - `NF_Name` receives the names of the sites.
   - `NF_Feature` class stores the identifier of each name together the `net:site` type identifier provide from `NF_Type` class.
   - `NF_Site` class receives the feature identifier (of a site, in this case) and its IP.
2. Population of the `NF_Name` and `NF_Feature` classes from NET:FCCN data. From this database are collected the domain names and the owner's postal code. All names of the domains are inserted into `NF_Name` class and its identifier into `NF_Feature` class together the identifier of the type `net:domain`.

Table 1: Distinct values from ADM:INE

Feature Type	#
NUT1	1
NUT2	7
NUT3	30
municipality	308

Table 2: Distinct values from POS:CTT

Feature Type	#
distrito	18
ilha	11
municipality	308
freguesia	3,595
localidade	44,386
zona	3,594
arruamento	146,422
código postal	187,014

4.3 Descriptive Statistics

In this section we show some of the results extracted from GKB for both domains, geographic and network. Results about the geo-administrative domain include both levels Portugal and World. These give the reader a quantitative information presently loaded in GKB.

4.3.1 Geo-administrative Domain

The results extracted from geographic domain include the data from ADM:INE, POS:CTT, ADM:WIKI and GEO:GAZ. Table 1 lists the number of features of each type provided from ADM:INE. Table 2 lists the number of features of each of the eight geographical types. It is possible to note that the number of *códigos postais* represents more than 50% (187,014) of the 266,212 distinct names (number of registers in `GF_Name` class) about Portugal inserted into GKB.

Other data included in GKB are the number of `municipalities` associated to *distritos*[1]. We present this information in Table 3 in decreasing order by number of `municipalities`.

The information about the population of each feature is used by some of the GKB applications for several purposes. For instance, one application uses the population value to disambiguate between features with identical names. Table 4 (a) presents the ten most populous `municipalities` and Table 4 (b) the ten less populous `municipalities` in Portugal according to the ADM:INE. The average of population by `municipalities` is 33,624 and the standard deviation is 54,870. This value of the standard deviation represents a large variation between the average and the value of the population of each `municipality`.

Table 5 presents descriptive statistics about GKB loaded with data about Portugal. Considering the all types of features in geographic domain, except postal code, there are 198,769 distinct names of geographic entities in GKB. In average, each name in GKB is associated to 2.5 features. This fact is an evidence that there is a lot of ambiguity in the geographic domain of Portugal.

Most of the relationships (99.83%) are of the `part of` type, while `equivalence` and `adjacency` are less frequent, since just `municipalities` and *localidades* have `equivalence` relation and just `municipalities` have `adjacency` relation.

Basically, each feature has a broader feature, while a feature has in average ten narrower features. Considering just the features with equivalents, we have about two equivalent features for each, while for the adjacent features this value increases to 3.54. For most of the features, there are no descendants, equivalent and adjacent, however just three features does not have

[1]Although the isles do not be considered properly *distritos* in ADM:INE, POS:CTT database considers them as *distrito*.

Table 3: Number of `municipalities` by *distrito*

Distrito/Ilha	# of `municipalities`
Viseu	24
Santarém	21
Aveiro	19
Porto	18
Coimbra	17
Faro	16
Leiria	16
Lisboa	16
Portalegre	15
Vila Real	14
Beja	14
Braga	14
Èvora	14
Guarda	14
Setúbal	13
Bragança	12
Castelo Branco	11
Viana do Castelo	10
Ilha da Madeira	10
Ilha de São Miguel	6
Ilha do Pico	3
Ilha de São Jorge	2
Ilha Terceira	2
Ilha das Flores	2
Ilha da Graciosa	1
Ilha de Porto Santo	1
Ilha de Santa Maria	1
Ilha do Corvo	1
Ilha do Faial	1

`part of` relation with other features. This fact points that most of the features have at least one relationship type connecting them.

4.3.2 Network Domain

The numbers presented in this section were extracted from NET:Versus and NET:FCCN information sources. Data provided by NET:FCCN were stored in GKB generating 39,191 domains. From these, there are 32,191 registers with at least one postal code associated under "PT" top level domain. These data are summarised in Table 6. We get to identify at least one *localidade* as a potential scope for each of the 32,191 domains under the "PT" top level domain.

Presently, we have 84,015 sites (73,278 from PT5 and 10,737 from PT4) provided from two crawls of the Portuguese Web by the tumba!.

Table 4: Population by `municipality`

a) Top ten

`Municipality`	Population
Lisboa	564,657
Sintra	363,749
Vila Nova de Gaia	288,749
Porto	263,131
Loures	199,059
Amadora	175,872
Cascais	170,683
Matosinhos	167,026
Braga	164,192
Gondomar	164,096

b) Lower ten

`Municipality`	Population
Corvo	425
Lajes das Flores	1,502
Barrancos	1,924
Santa Cruz das Flores	2,493
Alvito	2,688
Porto Moniz	2,927
Mourão	3,230
Vila de Rei	3,354
Arronches	3,389
Monforte	3,393

Table 5: Descriptive statistics of the Geographic Ontology of Portugal

Statistic	Value
Number of features	418,065
Number of distinct names different of postal codes	78,392
Number of features different of postal codes	198,769
Number of relationships	419,867
Number of part-of relationships	418,340 (99.83%)
Number of equivalence relationships	395 (0.09%)
Number of adjacency relationships	1,132 (0.27%)
Avg. broader features p/features	1.0016
Avg. narrower features p/features	10.5562
Avg. equivalent features p/features with equivalent	1.99
Avg. adjacent features p/features with adjacent	3.54
Number of features without ancestors	3 (0.00%)
Number of features without descendants	374,349 (89.54%)
Number of features without equivalent	417,867 (99.95%)
Number of features without adjacent	417,739 (99.92%)

Table 6: NET:FCCN Statistics

Property	Value
# of internet domains	39,191
# of Internet Domains with at least one valid postal code registered into Portugal	32,191
# of internet domains registered outside Portugal	3,012
# of distinct postal codes from Web domains	7,062
# of distinct postal codes from POS:CTT	187,014

5 Instance of the World

The GKB World repository is more than a simple aggregation of multilingual gazetteer lists. Names of continents, countries and administrative divisions among others, are inter-related

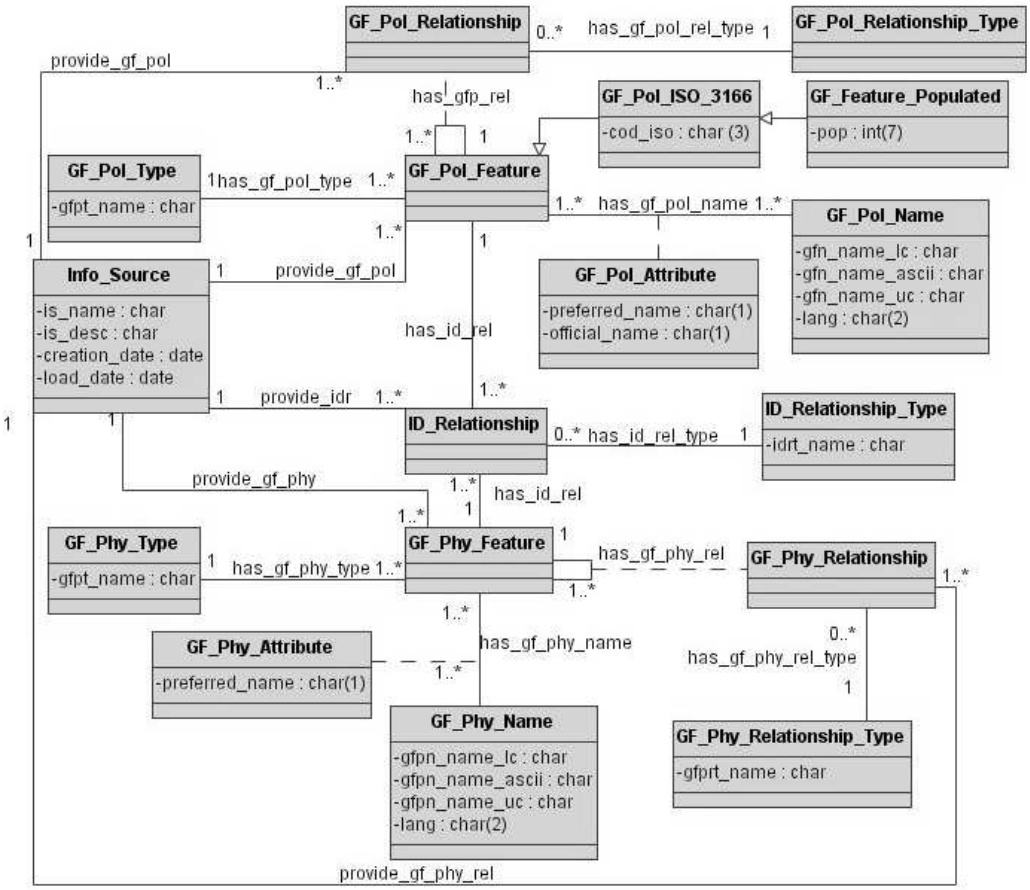

Figure 12: Full data model of GKB (instance of the World)

in a machine-readable way. We model two geographic domains in this instance of GKB: administrative and physical. Figure 12 presents the full data model of GKB with World data. Appendix C presents the SQL code to create relational schemas of the data models presented in this section.

Most of World data belong to the geo-administrative domain (classes with the prefix `GF_Pol` represent this domain). To capture all the World related information required by our applications, class `GF_Pol_Type` adds the types ISO 3166-1 (countries and territories), ISO 3166-2 (sub-entities into ISO 3166-1), city, place, agglomeration and administrative division as instances. Except for the ISO types, in this GKB instance only the geographic features with population above 100,000 people are stored. Countries and territories are usually referred both by their common (or short) name or official name. We capture this in two Boolean attributes, `preferred_name` and `official_name`. The `GF_Pol_Name` class stores the names of the features plus their language, which is maintained in the attribute `lang`.

The class `GF_Pol_Relationship_Type` the same relationships (part of and adjacency) defined to describe the geographic names on the GKB instance of Portugal. We also specialised the class `GF_Pol_Feature`. Class `GF_Pol_ISO3166` stores ISO3166 codes of countries, territories and

regions, while GF_Feature_Populated stores the population of the geographic features when available.

For the geo-physical domain, we defined the planet, continent, sea and lake. In class GF_Pol_Relationship, we related them. We are now introducing the names of the oceans and other geo-physical types. In class ID_Relationship, we relate the geo-administrative and geo-physical domains. All the countries and territories are related to their respective continents.

5.1 Information Sources

Gazetteer (W:GAZ): We obtain from the World gazetteer information about the largest cities and agglomerations around the world. An instance of W:GAZ is a record with information stating that the state of *Rio Grande do Sul* is located in Brazil, has population *10,723,745* and it is classified as an *administrative division*.

Wikipedia (ADM:WIKI): This is the same information source used to load the Portugal instance. However, for World data, we use the names of countries and capitals in four languages. In addition, we collect all the geo-physical domain from this source. An example from this source states that *Maputo* is the capital of *Mozambique*.

5.2 Loading Procedure

Loading of the geographic domain with World data involves the following steps:

1. Population of the Info_Source, GF_Pol_Type and GF_Pol_Relationship_Type classes.
2. Population of the GF_Pol_Name, GF_Pol_Feature and GF_Pol_Relationship classes with data about countries, territories and capitals in Portuguese, Spanish, English and German, respectively.
3. Population of the GF_Pol_Name, GF_Pol_Feature and GF_Pol_Relationship classes with data about ISO-3166-2, agglomeration, place and administrative division.
4. Population of the GF_Phy_Type and GF_Phy_Relationship_Type classes.
5. Population of the GF_Phy_Name, GF_Phy_Feature and GF_Phy_Relationship classes with data about planet, continents, seas and lakes.

5.3 Descriptive Statistics

Table 7 gives descriptive statistics of the GKB instance with World data. This is a smaller instance than one with geographic knowledge about Portugal. We show the detailed values about each feature type at the top of the table. The number of features is 12,283 and 7,970 of them have a population associated. Most of the features are provided by W:GAZ information source. Only relationship types `part of` and `adjacency` are used to connect all features. `Part of` relationships are the most common. It is worth mentioning that the features ISO-3166-1 contain preferred and alternative names, which includes the adjectives of the countries, such as Brazilian, Australian and Finnish. We load preferred and alternative names in four languages, while the adjectives are only defined for the English and German languages presently.

Table 7: Descriptive statistics of the Geographic Ontology of the World

Geographic Administrative Division	
Statistic	Value
Number of features ISO-3166-1 (4 languages)	239
Number of features ISO-3166-2 (in English)	3,979
Number of features Agglomeration (in English)	751
Number of features Place (in English)	3,968
Number of features Administrative Division (in English)	3,111
Number of features City-Capital (4 languages)	233
Number of features Regions (4 languages)	2
Number of features	**12,283**
Number of populated features	7,970 (64,88%)
Number of features from WIKI	4,453 (36,25%)
Number of features from W:GAZ	7,830 (63,75%)
Number of relationships part of	11,995
Number of relationships	**11,995**
Geographic Physical Division	
Statistic	Value
Number of features Planet (4 languages)	1
Number of features Continent (4 languages)	7
Number of features Sea (4 languages)	1
Number of features Lake (4 languages)	1
Number of features	**10**
Number of features from wikipedia	10 (100%)
Number of relationships part of	9
Number of relationships	**9**
Inter-Domain Relationships	
Statistic	Value
Number of relationships part of	241
Number of relationships adjacency	13
Number of relationships	**254**
Total	
Total number of features	**12,293**
Total number of relationships	**12,258**
Number of part-of relationships	12,245 (99,89%)
Number of equivalence relationships	2,501(20,40%)
Number of adjacency relationships	13 (0.10%)
Avg. broader features per feature	1.07
Avg. narrower features per feature	475.44
Avg. equivalent features per feature with equivalent	3.82
Avg. adjacent features per feature with adjacent	6.5
Number of features without ancestors	1(0.00%)
Number of features without descendants	12,045 (97,98%)
Number of features without equivalent	11,819 (96,14%)
Number of features without adjacent	12,291 (99,99%)

Figure 13: Components of GKB

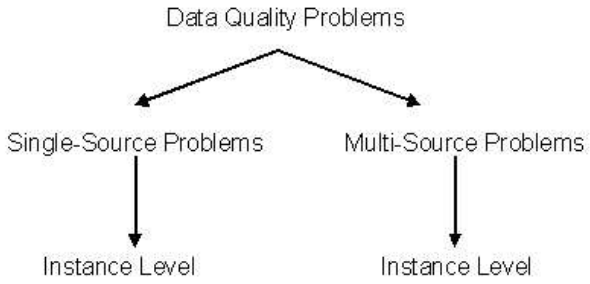

Figure 14: Data quality problems classification [Rahm and Do, 2000]

6 Data Quality and Information Integration Issues

In this section, we detail the process of data conversion and cleaning as it is loaded into GKB. In addition, we describe the subsequent data normalisation process. Finally, we discuss the semantic integration of the information collected from the GKB information sources.

The data sources used by GKB are independently developed and maintained to serve specific needs, resulting in 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 should be purged out and complementary information should be consolidated and merged in order to achieve a consistent view of the modeled world.

The process of data cleaning, which is essential to build a consistent KB, is abbreviately designed as ETL, after the initials of its three phases: extraction, transformation, loading [Rahm and Do, 2000]. A hard and rigorous work has to be done during this process, since the data which will be inserted into GKB comes from several sources and needs to be cleaned to eliminate inconsistences and duplicates.

ETL is carried out before data are inserted into GKB (see Figure 13). When the administrator receives the data to be inserted into GKB, he stores them into a set of files in the Comma Separated Values (CSV) format. After this, he implements some scripts or modifies those already available to perform the data cleaning. In the sequence, the data are loaded into the tables of GKB and other scripts are run to generate the domain ontologies in OWL standard.

Rahm and Do created a classification of data cleaning problems (depicted in Figure 14) that splits the data quality problems in single-source and multi-source problems [Rahm and Do, 2000]. In our work, we faced problems of both kinds, which are elaborated on the remainder of this section.

Table 8: Examples of spelling errors

ID	Name of the feature
193771	bairro da cooperariva 1ª fase
193772	bairro da coopertiva 1ª fase
193773	bairro da cooperativa 1ª fase
193774	bairro da cooperativa 1º fase

Table 9: Inconsistences in postal code from network domains database

Domain	postal code
mestredeaviz	1495 -148
centralfundos	1050 - 185
sos	1600162
adruse	6290–520
hotelinfantesagres	4050-
belo-construcoes	1070
moviflor	199-008
esteproar	2.735.507
fetec	800

6.1 Single-source Problems

The main tasks in solving these problems include correction of the spelling errors, validation and correction of postal codes, insertion of alternative names, and correction of geographic coordinates. Some of the handled cases found while cleaning the data used to build the GKB instance of Portugal are detailed below.

6.1.1 Spelling Errors Correction

Spelling errors are common in large information sources due to the data having been generated by humans. There are multiple cases where one name is identified by more than one identifier due to spelling errors (see the examples in Table 8). In the GEO:GAZ database there are also other spelling errors. We found the location *Estoril* spelled as *estroil*. In this case, the letters *o* and *r* were transposed. In GKB, such errors are detected and eliminated whenever found by the repository administrator. However, the deletion of all occurrences of spelling errors is an exhaustive, tedious and slow task. So, it still possible to find errors of this kind in GKB.

6.1.2 Postal Codes Validation and Correction

We have found some invalid postal codes in the network domains database when trying to match them with the postal data from POS:CTT. Analysing some examples, we can detect cases where it is possible to validate some of them, as shown in Table 9.

To detect some of the incorrect postal codes like those presented in Table 9, we implemented scripts that do:

- identify the sequence of postal code digits;
- normalise the postal code representation;
- validate the postal codes against the CTT ones.

Table 10: Coordinates to the region of *Vila Nova* at *distrito* of *Viseu*

Latitude	Longitude
40.4167	-8.2167
41.0833	-8.1167
41.0833	-8.0333
40.6500	-8.0167
40.9000	-8.0000
40.7000	-7.9333
41.0500	-7.8833
41.0500	-7.6833

Besides this validation, we implemented a process for normalisation for the names provided from multiple sources. It includes the removal of extra white spaces and invalid characters from strings, elimination of carriage return in the end of strings and conversion to lower case.

6.1.3 Insertion of Alternative Names

We also store alternative names in GKB in order to help the future process of query expansion. Data from GEO:GAZ contain words with cedilla and accented characters together with alternative names having the accents. In this case, we match just the data with accented characters. Other names are stored as alternative names with the relation `equivalent to` to the preferred name.

We also found alternative place names to some regions in the GEO:GAZ. For example, São João, located in *distrito Viana do Castelo*, has the following alternative names: *Vila Chã* and *São João Baptista*.

6.1.4 Correction of Geographic Coordinates

Sometimes a region is associated with different coordinates. For example, in GEO:GAZ *Vila Nova*, located in *distrito* of *Viseu*, has eight different coordinates (shown in Table 10).

When one region has more than one geographic coordinate, we calculate the average of the values and store just it. In GKB, the region of *Vila Nova* at *distrito* of *Viseu* has latitude $40.8667^o N$ and the longitude $7.9854^o W$.

6.2 Multi-source Problems

Besides the problems found in single-sources, we find inconsistences when matching data from distinct geographic information sources and geographic data with network data. We describe both in the following.

6.2.1 Matching between Data from POS:CTT and GEO:GAZ

When matching data from POS:CTT and GEO:GAZ, we found nine *distritos* POS:CTT (*Ilha Terceira, Ilha da Graciosa, Ilha das Flores, Ilha de Santa Maria, Ilha de São Jorge, Ilha de São Miguel, Ilha do Corvo, Ilha do Faial and Ilha do Pico*) that are present in the GEO:GAZ as *Região Autónoma dos Açores*. All these isles constitute this region. So, when matching each occurrence of *Região Autónoma dos Açores*, we verify if it is possible to find some of the isles cited before. In other words, to solve this mismatch, we replace names in GEO:GAZ by the corresponding names stored in POS:CTT.

Table 11: Types of geographic features

Domain	Feature Type 1	Relation	Feature Type 2
Geographic	NUT2	part of	NUT1
	NUT3	part of	NUT2
	`municipality`	part of	NUT3
	ilha	part of	NUT3
	`municipality`	part of	*ilha*
	`municipality`	part of	*distrito*
	localidade	part of	`municipality`
	freguesia	part of	`municipality`
	zona	part of	*localidade*
	arruamento	part of	*localidade*
	postal code	part of	*localidade*
Inter-Domain	domain	hasScope	*localidade*

Table 12: Inconsistences between information sources and Web data

Internet Domains		Postal Data		Data from Web site	
Domain	PC	*rua* and/or *localidade*	PC	Address from site	PC
atlanticopress	3810-185	Rua São Martinho - Aveiro	3810-185	Av. Luis Bivar,73 1^{o}Dt, Lisboa	1050-142
cm-lisboa	1100-060	Rua Áurea, Ímpares de 27 a 151 - Lisboa	1100-060	Praça do Município - Lisboa	1100-365

After the cleaning phase, we must load the cleaned data into the tables in order to allow us perform queries to generate the ontologies to the consumers.

The types of features used in GKB are associated through the relations depicted in Table 11. The graphical representation of this types, which constitute our geographic ontology, is given in Section 6.4, Figure 17.

6.2.2 Inconsistences between Information Sources and Web Data

Postal codes from Web domains are those that registrants provided. In some cases, these are not the postal codes appearing in the corresponding Web sites. Consequently, we should have in mind that this information can induce us to attribute an incorrect scope to a Web domain. Table 12 shows some examples where the same postal codes from information sources do not match with the postal codes found in the Web site.

In face of these inconsistences, we choose to associate geographic scopes to the features of type *localidade* that contain the *arruamento* that correspond to the postal code.

We also found the inconsistences between the data from the same information source when doing the update of GKB. For example, the `municipality` *vila velha de rodão* in the first file received from ADM:INE is without acute accent, while the same `municipality` *vila velha de ródão* has the accent in the version to update.

6.3 Data Normalisation

The cleaned names inserted into GKB are lowercase. However, these names should be correctly spelled in the ontologies, with the first letter of each word, except prepositions, capitalised. Although, this task seems initially simple and easy, we found some hard cases, as described in the following:

Roman numbers: Characters identified as roman numbers should be capitalised. We solve this problem by defining two hash tables: one of the roman characters and another with exceptions words (i.e., *civil*). We capitalise all names that are in the former but not in the latter.

Prepositions: Some articles assign the name grammatical category and should be capitalised, for example *entre* in *Entre-Campos*. We always capitalise the preposition if it appears in the beginning of the name. However, in the names as *Entre As Ruas Alfredo Pinto e Alfredo Feio*, the word *Entre* is a preposition. We do not have any tool to help in this task, so it is still possible to find capitalised prepositions.

Articles: Some vowels should be capitalised such as *O* in *Jornal O Povo de Cortegaça*. We capitalise articles when they are preceded by the words *Jornal* and *Revista* and at the beginning of a name (i.e., *O Algarve*).

Special characters: Quotes, parentheses and other special characters should be considered when we capitalise words. We capitalise the first letter of the word after the special character, once these kinds of characters are never followed by prepositions in our database.

Apostrophes: Occasionally, some names started the letter `d` followed by the apostrophe. In these cases, `d` is not capitalised, but the first character after it is capitalised.

Dots: When a word is composed by just one letter and this letter is followed by a dot, we capitalise it. This rule is used automatically to capitalise the acronyms that are typed with dots, (i.e., A.E.P.).

Acronyms: When they are typed without dot separating the letters, we do not identify them. To solve this problem, we create a hash table with the most common acronyms as (*CP*, *CTT*, *EDP*). Words in this hash table are always capitalised. However, this solution is not exhaustive.

We try to solve most of the problems concerning the capitalisation of the names of the geographic features. Although our methods works well to the cases described above, we are aware that it is still possible find lower case characters instead of upper case and vice-versa.

6.4 Semantic Integration

GKB receives information from multiple sources, each one with knowledge organised 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 15 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 15): at first, it searches the lowest common features types in both hierarchies (municipality). If it holds, it identifies the common instances between the hierarchies (*Matosinhos*, *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

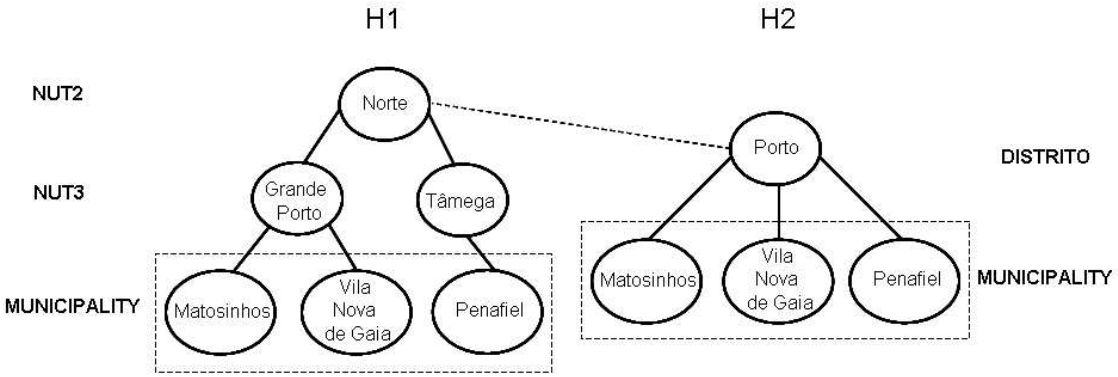

Figure 15: Feature types dependencies in two information sources of the GKB instance of Portugal

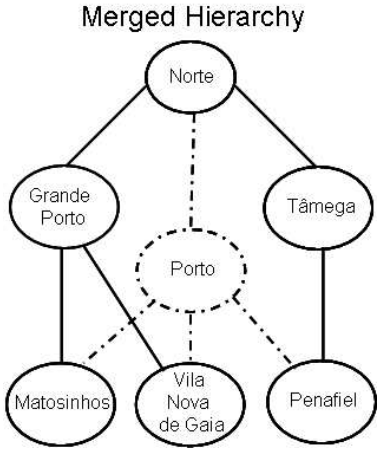

Figure 16: Merged GKB hierarchy

H2). After these steps, the algorithm verifies the distance (in number of relationships `part of`) 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 `part of` with the ancestor (`Norte`) in the another hierarchy. The existing relationships in both hierarchies are maintained. Figure 16 shows the merged hierarchy.

Figure 17 represents the merged hierarchies of all the information sources used to load the GKB instance of Portugal obtained with the process described above.

7 Representing Geographic Knowledge in GKB

GKB not only manages geographic and geographic-related entities and relationships, but also the rules relating them. New knowledge is incorporated in GKB as rules. 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

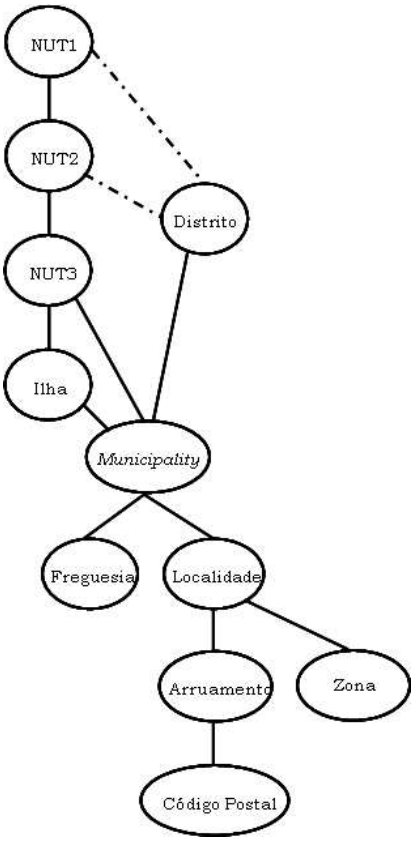

Figure 17: Graphical representation of the feature types dependencies in the geo-administrative domain of the GKB instance of Portugal

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 18 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, 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 the network domain, we represent the URL of sites tokenised in three atomic concepts: subdomain, 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

```
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 18: 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)

Table 13: Rule-based assigned scopes by GKB to sites of Portugal

Site Type	# of sites	# of unifications	Site Type	# of sites	# of unifications
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%)

`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): For instance, 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)` $\sqcap$
(`∃netSitePrefix(idN,‘‘cm’’)` $\sqcup$ `∃netSitePrefix(idN,‘‘mun’’)`) $\sqcap$
`∃geoFeatureType(idG,‘‘CON’’)` $\sqcap$ `∃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 an unification 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`.

We could assign scopes to most of the sites in GKB instances of Portugal unifying the rules above. However, these unifications 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`.

Table 13 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 unifications obtained 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.

8 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 GKB instance.

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

8.1 Declaring the Vocabularies to Be Used in the Geographic Ontology

In the ontologies generated, we need to indicate the vocabularies used at the beginning of their descriptions. These vocabularies are described through a set of XML namespaces declarations as follows:

```
<rdf:RDF
  xmlns:gn = "http://xldb.di.fc.ul.pt/geo_net_pt01.owl#"
  xmlns:owl = "http://www.w3.org/2002/07/owl#"
  xmlns:rdf = "http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:rdfs= "http://www.w3.org/2000/01/rdf-schema#"
  xmlns:xsd = "http://www.w3.org/2001/XMLSchema#">
```

The first two declarations identify the namespace associated with this ontology. The first makes it the *default* namespace, stating that unprefixed qualified names refer to the current ontology. The second identifies the namespace of the current ontology with the prefix *gn:*. The third declares that in this document, elements prefixed with *owl:* should be understood as referring to things drawn from the namespace (`http://www.w3.org/2002/07/owl`). The last three namespaces refer to RDF, RDFS and XML Schema datatypes (more details about these standards can be found in *http://www.w3.org*), respectively.

After we declare the namespaces, information about the ontology are inserted under the *owl:Ontology* tag. The content of this tag is important to housekeeping tasks.

```
<owl:Ontology rdf:about="GKB\_Ontology">
  <rdfs:comment>Description of classes and properties of Geographic Ontology
  of Portugal</rdfs:comment>
  <owl:priorVersion rdf:resource="http://www.ourmachine.pt/gkb.owl/20050630-gkb.owl"/>
  <rdfs:label>Grease Ontology</rdfs:label>
</owl:Ontology>
```

The *rdf:about* attribute provides a name for the ontology, while the *rdfs:comment* tag gives an overview about the ontology been described. In the following, the *owl:priorVersion* tag points to the URL where is this ontology. The *rdfs:label* tag helps Web agents find the Geographic Ontology of Portugal in the internet.

In GKB ontology, each feature has an unique identifier, a name, a type, a footprint composed by a latitude and a longitude when the type is a *localidade* and a relationship with other features. This relationship can be twofold: `part of` or `equivalent`. The former declares the meronymy semantic relation, while the latter indicates the available alternative names.

Appendix C presents the full GKB ontology in OWL, which was validated in (`http://www.w3.org/RDF/Validator/` and `http://phoebus.cs.man.ac.uk:9999/OWL/Validator`).

8.2 GOG - GKB Ontology Generator

GOG enables the generation of the instances stored in GKB in a suitable format requested by users. This format is established in a vocabulary apart from GOG.

Figure 19 presents an excerpt of the representation of an instance of GKB with data about Portugal as an ontology. The excerpt describes the feature type `municipality` (abbreviated as `CON`) named *Porto*, which has identifier *GEO_238*. This feature was imported from *Instituto*

```
<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>
```

Figure 19: An excerpt of GKB-extracted ontology with data about Portugal

Nacional de Estatística (`INE`). The `municipality` 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 `municipality` of *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 it is available as a public resource in `http://xldb.di.fc.ul.pt/geonetpt`.

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 sources directly available on the Web. Figure 20 presents an excerpt of this ontology, which has more than 14,000 features.

It shows the description of the feature *Germany*, which is identified by *GEO_9* and its type is `ISO-3166-1`. Germany has preferred (represented by the *geo_name* tag) and alternative names in English, Portuguese, Spanish and German. It has two relationships: the former, *part-of* (`PRT`) the feature *Phy_7* (*Europe*) and the latter, *adjacency* (`ADJ`) of the feature *Phy_9* (*North Sea*). Both *Europe* and *North* Sea are declared in another part of this same ontology.

This information is provided from `WIKI`, the identifier to the information source wikipedia. Appendix D gives the full vocabulary of the Geographic World Ontology.

```
<gn:Geo_Feature rdf:ID="GEO_9">
  <gn:geo_id>9</gn:geo_id>
  <gn:geo_name xml:lang="en">Germany</gn:geo_name>
  <gn:alternative_name>
    <rdf:Bag>
      <rdf:li>
        <gn:Geo_Name>
          <gn:geo_name xml:lang="de">Alemanha</gn:geo_name>
        </gn:Geo_Name>
      </rdf:li>
      <rdf:li>
        <gn:Geo_Name>
          <gn:geo_name xml:lang="es">Alemania</gn:geo_name>
        </gn:Geo_Name>
      </rdf:li>
      <rdf:li>
        <gn:Geo_Name>
          <gn:geo_name xml:lang="de">Deutschland</gn:geo_name>
        </gn:Geo_Name>
      </rdf:li>
      <rdf:li>
        <gn:Geo_Name>
           <gn:geo_name xml:lang="pt">República Federal da Alemanha</gn:geo_name>
        </gn:Geo_Name>
      </rdf:li>
      <rdf:li>
        <gn:Geo_Name>
           <gn:geo_name xml:lang="es">República Federal de Alemania</gn:geo_name>
        </gn:Geo_Name>
      </rdf:li>
      <rdf:li>
        <gn:Geo_Name>
           <gn:geo_name xml:lang="de">die Bundesrepublik Deutschland</gn:geo_name>
        </gn:Geo_Name>
      </rdf:li>
      <rdf:li>
        <gn:Geo_Name>
           <gn:geo_name xml:lang="en">Federal Republic of Germany</gn:geo_name>
        </gn:Geo_Name>
      </rdf:li>
      <rdf:li>
        <gn:Geo_Name>
           <gn:geo_name xml:lang="en">German</gn:geo_name>
        </gn:Geo_Name>
      </rdf:li>
      <rdf:li>
        <gn:Geo_Name>
           <gn:geo_name xml:lang="de">deutsch</gn:geo_name>
        </gn:Geo_Name>
      </rdf:li>
    </rdf:Bag>
  </gn:alternative_name>
  <gn:geo_type_id rdf:resource="http://xldb.di.fc.ul.pt/geo-net.owl#ISO-3166-1"/>
  <gn:related_to>
      <rdf:Bag>
         <rdf:li>
           <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_PHY_7"/>
           </gn:Geo_Relationship>
         </rdf:li>
         <rdf:li>
           <gn:Geo_Relationship>
             <gn:rel_type_id rdf:resource="http://xldb.di.fc.ul.pt/geo-net.owl#ADJ"/>
             <gn:geo_id rdf:resource="#GEO_PHY_9"/>
           </gn:Geo_Relationship>
         </rdf:li>
      </rdf:Bag>
  </gn:related_to>
  <gn:info_source_id rdf:resource="http://xldb.di.fc.ul.pt/geo-net.owl#WIKI"/>
</gn:Geo_Feature>
```

Figure 20: An excerpt of an ontology extracted from GKB repository with World data

9 Applications using GKB

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 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, Chen et al., 1998]. 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 capitalisation and the surrounding text. Mikheev et al. showed that 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. Cucchiarelli et al report that one of the bottlenecks in designing NER systems is the limited availability of large gazetteers [Cucchiarelli et al., 1998]. Our NER system for geographical names uses the information at GKB as the main dataset, together with some simple hand-coded rules [Martins and Silva, a, Martins and Silva, c]. 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 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. Recognising geographical named entities in a document is also in itself not enough for classification, as geographical entities are ambiguous [Page et al., 1999, Sang et al., 2003]. We developed a specific method for this problem that instead of the standard machine learning methodology of automatically inferring classifiers from a training set of documents uses the recognised geographical named entities together with a combination/disambiguation algorithm that builds on the GKB ontological relationships [Martins and Silva, b]. The disambiguation algorithm sees the ontology as a graph and takes its inspiration on PageRank [Baeza-Yates and Davis, 2004, Mihalcea and Tarau, 2004]. The geographical 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 provides the information used to present users with different alternatives to their queries. Figure 21 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. In the background, Geo-Tumba uses the GKB to attribute a scope to the Web sites. When an ambiguous geographic name is detected in the query, Geo-Tumba shows possible alternatives to user disambiguates its query. For example, the name *rua Castelo Branco* occurs in five different `municipalities`, which are presented in the left inferior side of the Figure 21. Further the text query, the user can use maps to define the scope of a query.

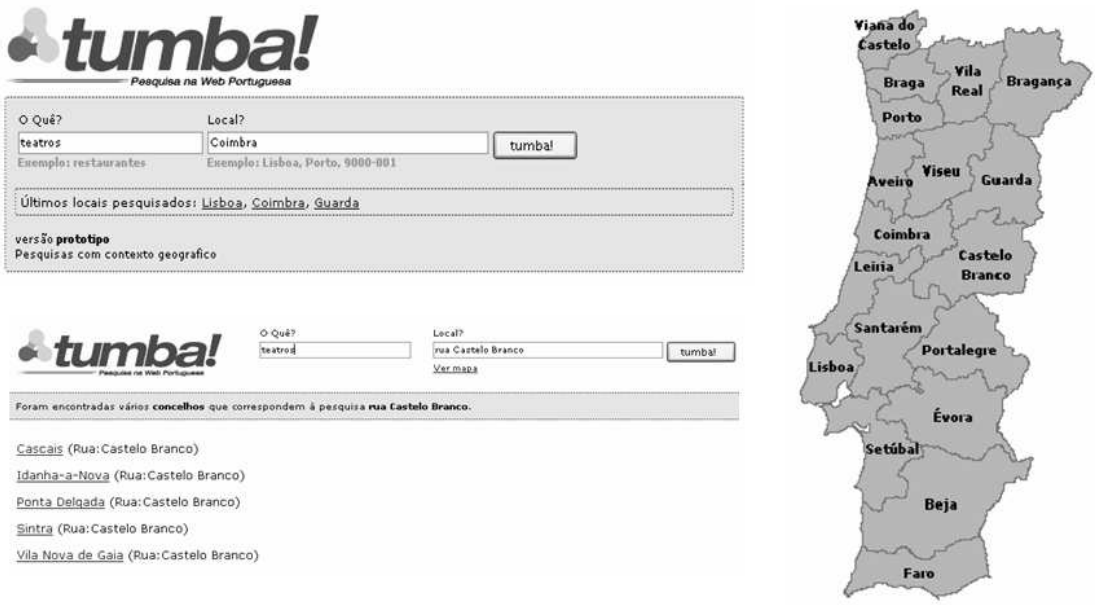

Figure 21: Interface of the geographic search engine using GKB

We used 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 was 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.

10 Final Remarks

We presented GKB, a repository based on a domain-independent meta-model for integrating geographic knowledge collected from multiple sources. We gave an overview of GKB through its context, requirements and architecture, which is composed by information domains. We detailed the instances of Portugal and World data. Next, we described the pre-processing phase to enhance the quality of data before it is loaded in GKB. Most of the inconsistences are eliminated in this phase.

Once GKB has been loaded with data from its sources, several information integration issues remain to be addressed. Some can be solved by using geographic knowledge, which allows assigning inter-domain relationships and geographic scopes to Web sites.

We also presented the ontologies generated from GKB for both Portugal and World instances. The ontologies represent an uniform vision of the previously distributed information. The content of other databases can be migrated to the Semantic Web using some of these ontologies. Finally the applications using GKB were described. GKB could be used to manage similar knowledge from any other country or region and serve as repository for other applications than those which we have developed.

We are in the process of augmenting the knowledge present in this repository with the semantic relations between the geographic entities extracted from the texts of the Portuguese Web. We will use of the semantic relations identified in GKB plus natural language processing techniques to aid the identification of the other geographic relations in Web texts. This process should be iterative in the next years, expanding the existing 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 and all the GREASE participants.

References

[Alani et al., 2003] 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.

[Baader et al., 2003] 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 and Davis, 2004] 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.

[Berners-Lee et al., 2001] Berners-Lee, T., Hendler, J., and Lassila, O. (Issue 01/05/2001, 2001). The Semantic Web. A New Form of Web Content that is Meaningful to Computers will Unleash a Revolution of New Possibilities. *Scientific American*.

[Chen et al., 1998] Chen, H., Ding, Y., and Tsai, S. (1998). Named entity extraction for information retrieval. *Computer Processing of Oriental Languages, Special Issue on Information Retrieval on Oriental Languages*, 12(1):75–85.

[Cruz et al., 2002] Cruz, I. F., Rajendran, A., Sunna, W., and Wiegand, N. (2002). Handling semantic heterogeneities using declarative agreements. In *Proc. of the 10th ACM International Symposium on Advances in Geographic Information Systems - ACM-GIS 2002*, pages 168–174.

[Cucchiarelli et al., 1998] Cucchiarelli, A., Luzi, D., and Velardi, P. (1998). Automatic semantic tagging of unknown proper names. In *Proceedings of the 17th international conference on Computational linguistics*, pages 286–292, Morristown, NJ, USA. Association for Computational Linguistics.

[Fensel, 2001] Fensel, D. (2001). *Ontologies: Silver Bullet for Knowledge Management and Electronic Commerce*. Springer-Verlag.

[Fu et al., 2003] Fu, G., Abdelmoty, A. I., and Jones, C. (2003). Design of a Geographical Ontology. Technical report, D5 3101 - SPIRIT - Spatially-Aware Information Retrieval on the Internet.

[Gomes et al., 2002] Gomes, D., Campos, J. P., and Silva, M. J. (2002). Versus: a Web Repository. In *WDAS - Workshop on Distributed Data and Structures 2002*, Paris, France.

[Gonzalez, 2001] Gonzalez, M. (2001). Thesauri. Trabalho individual, Faculdade de Informática - PUCRS.

[Gravano et al., 2003] Gravano, L., Hatzivassiloglou, V., and Lichtenstein, R. (2003). Categorizing web queries according to geographical locality. In *12th ACM Conference on Information and Knowledge Management (CIKM 2003) - New Orleans, Louisiana, USA*, pages 325 – 333, New York, NY, USA. ACM Press.

[Gruber, 1993] Gruber, T. (1993). A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition*, 5(2):199–220.

[Guarino, 1997] Guarino, N. (1997). Understanding, Building and Using Ontologies. A commentary to "Using Explicit Ontologies in KBS Development". *International Journal of Human and Computer Studies*, 46:293–310.

[Hill, 2000] Hill, L. L. (2000). Core elements of digital gazetteers: Placenames, categories, and footprints. In *In Research and Advanced Technology for Digital Libraries: Proc. of the 4th European Conference, ECDL2000*, pages 280–290.

[Hyvönen et al., 2004] Hyvönen, E., Salminen, M., Junnila, M., and Kettula, S. (2004). A content creation process for the semantic web. In Oltramari, A., Paggiio, P., Gangemi, A., Pazienza, M. T., Calzolari, N., Pedersen, B. S., and Simov, K., editors, *Workshop Ontolex - Ontologies and Lexical Resources in Distributed Environments*, pages 9–15.

[Inoue et al., 2002] Inoue, Y., Lee, R., Takakura, H., and Kambayashi, Y. (2002). Web locality based ranking utilizing location names and link structure. In *Second International Workshop on Web and Wireless Geographical Information Systems (W2GIS 2002)*, pages 56–63. IEEE.

[Irie and Sundheim, 2004] 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] ISO19109 (2005). ISO 19109. https://www.seegrid.csiro.au/twiki/pub/Xmml/FeatureModel/19109_DIS2002.pdf.

[ISO2788, 1986] ISO2788 (1986). International Organization for Standardization - Documentation - Guidelines for the establishment and development of monolingual thesauri. Geneva.

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

[Kietz et al., 2000] Kietz, J., Maedche, A., and Volz, R. (2000). A Method for Semi-Automatic Ontology Acquisition from a Corporate Intranet. In *12th International Conference on Knowledge Engineering and Knowledge Management EKAW'2000 - Workshop on Ontologies and Texts*, Juan-les-Pins,French Riviera.

[Manov et al., 2003] 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*.

[Markowetz et al., 2004] Markowetz, A., Brinkhoff, T., and Seeger, B. (2004). Geographic Information Retrieval. In *3rd International Workshop on Web Dynamics*.

[Martins and Silva, a] Martins, B. and Silva, M. J. Categorizing web pages according to geographical scopes. (To Appear).

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

[Martins and Silva, c] Martins, B. and Silva, M. J. The webcat framework : Automatic generation of meta-data for web resources. (To Appear).

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

[Mena, 1998] Mena, E. (1998). *OBSERVER: An Approach for Query Processing in Global Information Systems based on Interoperation across Pre-existing Ontologies*. PhD thesis, Departamento de Informática e Ingeniería de Sistemas. Universidad de Zaragoza.

[Mihalcea and Tarau, 2004] Mihalcea, R. and Tarau, P. (2004). TextRank: Bringing order into texts. In *Proceedings of EMNLP-04, the 2004 Conference on Empirical Methods in Natural Language Processing*.

[Mikheev et al., 1999] 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.

[Nobécourt, 2000] Nobécourt, J. (2000). A method to build formal ontologies from texts. In *12th International Conference on Knowledge Engineering and Knowledge Management*, Juan-les-Pins,French Riviera.

[Page et al., 1999] 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.

[Purves and Jones, 2004] Purves, R. and Jones, C. (2004). Workshop on Geographic Information Retrieval, SIGIR 2004. Disponível em: http://www.geo.unizh.ch/∼rsp/gir/program.html.

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

[Sang et al., 2003] Sang, T. K., Erik, F., and Meulder, F. D. (2003). Introduction to the CoNLL-2003 shared task: Language-Independent Named Entity Recognition. In Daelemans, W. and Osborne, M., editors, *Proceedings of CoNLL-2003, the 7th Conference on Natural Language Learning*, pages 142–147. Edmonton, Canada.

[Sheth et al., 2004] 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*.

[Szulman et al., 2002] Szulman, S., Biébow, B., and Aussenac-Gilles, N. (2002). Structuration de Terminologies à l'aide d'outils de TAL avec TERMINAE. *Revue Traitement Automatique des Langues*, 43(1).

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

Appendix A - Geo-administrative feature types defined in GKB (instance of Portugal)

This appendix lists the geo-administrative feature types exiting in Portugal. The list is composed by the identifier of the feature type followed by its full name. The identifiers are used in both, GKB and ontologies generated from GKB.

Identifier	Geographic Feature Type
ACE	adm:arruamento:acesso
ADR	adm:arruamento:adro
ALA	adm:arruamento:alameda
AVE	adm:arruamento:avenida
AZI	adm:arruamento:azinhaga
BAI	adm:arruamento:bairro
BEC	adm:arruamento:beco
CAI	adm:arruamento:cais
CAL	adm:arruamento:calçada
CAM	adm:arruamento:caminho
CAR	adm:arruamento:carreira
CDP	adm:código postal
CON	adm:concelho
CPO	adm:arruamento:campo
CTO	adm:arruamento:canto
DST	adm:distrito
ENH	adm:arruamento:escadinhas
ESC	adm:arruamento:escadas
EST	adm:arruamento:estrada
FRG	adm:freguesia
JAR	adm:arruamento:jardim
LAD	adm:arruamento:ladeira
LOC	adm:localidade
LOT	adm:arruamento:loteamento
LRG	adm:arruamento:largo
LUG	adm:arruamento:lugar
MON	adm:arruamento:monte
NT1	adm:NUT1

Identifier	Geographic Feature Type
NT2	adm:NUT2
NT3	adm:NUT3
OUT	adm:arruamento:outro
PAI	adm:país
PAR	adm:arruamento:parque
PAS	adm:arruamento:passeio
PAT	adm:arruamento:pátio
PRA	adm:arruamento:praça
PTA	adm:arruamento:praceta
PTE	adm:arruamento:ponte
QUE	adm:arruamento:quelha
QUI	adm:arruamento:quinta
REC	adm:arruamento:recanto
RLA	adm:arruamento:ruela
ROT	adm:arruamento:rotunda
RPA	adm:arruamento:rampa
RUA	adm:arruamento:rua
SIT	adm:arruamento:sítio
TER	adm:arruamento:terreiro
TRV	adm:arruamento:travessa
URB	adm:arruamento:urbanização
VER	adm:arruamento:vereda
VIA	adm:arruamento:via
VIE	adm:arruamento:viela
VLE	adm:arruamento:vale
ZNA	adm:arruamento:zona
ZON	adm:zona

Appendix B - SQL script to create a relational schema of GKB instance of Portugal

This appendix presents a SQL script to create a relational schema to GKB instance of Portugal.

```
# Table of Information Source, used by both Geo and Net Domains
CREATE TABLE Info_Source (
  is_id INT(3) NOT NULL,
  is_name VARCHAR(255) NOT NULL,
  is_desc VARCHAR(255),
  creation_date date NOT NULL,
  load_date date NOT NULL,
  PRIMARY KEY (is_id,load_date)
);

# Tables of the Geo Domain
CREATE TABLE GF_Name (
  gfn_id INT UNSIGNED NOT NULL PRIMARY KEY,
  gfn_name VARCHAR(255) NOT NULL,
);

CREATE TABLE GF_Type (
  gft_id char(3) PRIMARY KEY,
  gft_name VARCHAR(255) NOT NULL,
  gft_desc VARCHAR(255)
);

CREATE TABLE GF_Feature (
  gff_id INT UNSIGNED NOT NULL PRIMARY KEY,
  gft_id INT NOT NULL REFERENCES GF_Type (gft_id) ON DELETE CASCADE,
  gfn_id INT UNSIGNED NOT NULL REFERENCES GF_Name (gfn_id) ON DELETE CASCADE,
  is_id INT(3) NOT NULL REFERENCES Info_Source (is_id) ON DELETE CASCADE,
  preferred_name char(1)
);

CREATE TABLE GF_Relationship_Type (
  gfrt_id CHAR(3) NOT NULL PRIMARY KEY,
  gfrt_name VARCHAR(255) NOT NULL,
  gfrt_desc VARCHAR(255)
);

CREATE TABLE GF_Relationship (
  gfr_id INT UNSIGNED NOT NULL,
  gff_id1 INT UNSIGNED NOT NULL REFERENCES GF_Feature (gff_id) ON DELETE CASCADE,
  gff_id2 INT UNSIGNED NOT NULL REFERENCES GF_Feature (gff_id) ON DELETE CASCADE,
  gfrt_id CHAR(3) NOT NULL REFERENCES GF_Relationship_Type(gfrt_id) ON DELETE CASCADE,
  is_id INT(3) NOT NULL REFERENCES Info_Source (is_id) ON DELETE CASCADE,
  PRIMARY KEY(gff_id1,gff_id2)
);

CREATE TABLE GF_Footprint (
  gffp_id INT UNSIGNED NOT NULL PRIMARY KEY,
  gff_id INT UNSIGNED REFERENCES GF_Feature (gff_id) ON DELETE CASCADE,
  latitude decimal(6,4) NOT NULL,
  longitude decimal(6,4) NOT NULL
);

CREATE TABLE GF_Feature_Populated (
  gffp_id INT UNSIGNED NOT NULL PRIMARY KEY,
  gff_id INT UNSIGNED NOT NULL REFERENCES GF_Feature (gff_id) ON DELETE CASCADE,
  pop INT NOT NULL
```

```
);


# Tables of the Net Domain
CREATE TABLE NF_Name (
  nfn_id INT UNSIGNED NOT NULL PRIMARY KEY,
  nfn_name VARCHAR(255) NOT NULL
);

CREATE TABLE NF_Type (
  nft_id char(3) PRIMARY KEY,
  nft_name VARCHAR(255) NOT NULL,
  nft_desc VARCHAR(255)
);

CREATE TABLE NF_Feature (
  nff_id INT UNSIGNED NOT NULL PRIMARY KEY,
  nft_id INT NOT NULL REFERENCES NF_Type (nft_id) ON DELETE CASCADE,
  nfn_id INT UNSIGNED NOT NULL REFERENCES NF_Name (nfn_id) ON DELETE CASCADE,
  is_id INT(3) NOT NULL REFERENCES Info_Source (is_id) ON DELETE CASCADE
);

CREATE TABLE NF_Relationship_Type (
  nfrt_id CHAR(3) NOT NULL PRIMARY KEY,
  nfrt_name VARCHAR(255) NOT NULL,
  nfrt_desc VARCHAR(255)
);

CREATE TABLE NF_Relationship (
  nfr_id INT UNSIGNED NOT NULL,
  nf_id1 INT UNSIGNED NOT NULL REFERENCES NF_Feature (nff_id) ON DELETE CASCADE,
  nf_id2 INT UNSIGNED NOT NULL REFERENCES NF_Feature (nff_id) ON DELETE CASCADE,
  type_rel CHAR(3) NOT NULL REFERENCES NF_Relationship_Type(nfrt_id) ON DELETE CASCADE,
  is_id INT(3) NOT NULL REFERENCES Info_Source (is_id) ON DELETE CASCADE,
  PRIMARY KEY(nf_id1,nf_id2)
);

CREATE TABLE NF_Domain (
  nfd_id INT UNSIGNED NOT NULL PRIMARY KEY,
  nff_id INT UNSIGNED NOT NULL REFERENCES NF_Feature (nff_id) ON DELETE CASCADE,
  owner_postal_code VARCHAR(255) NOT NULL
);

CREATE TABLE NF_Site (
  nfs_id INT UNSIGNED NOT NULL PRIMARY KEY,
  nff_id INT UNSIGNED NOT NULL REFERENCES NF_Feature (nff_id) ON DELETE CASCADE,
  ip VARCHAR(15) NOT NULL
);
#Tables of the inter-domain relationships
CREATE TABLE ID_Relationship_Type (
  idrt_id CHAR(3) NOT NULL PRIMARY KEY,
  idrt_name VARCHAR(255) NOT NULL,
  idrt_desc VARCHAR(255)
);
CREATE TABLE ID_Relationship (
  idr_id INT UNSIGNED NOT NULL,
  id_id1 INT UNSIGNED NOT NULL,
  id_id2 INT UNSIGNED NOT NULL,
  type_rel CHAR(3) NOT NULL,
  is_id INT(3) NOT NULL REFERENCES Info_Source (is_id) ON DELETE CASCADE,
  PRIMARY KEY(id_id1,id_id2,is_id));
```

Appendix C - SQL script to create a relational schema of GKB instance of the World

This appendix presents a SQL script to create a relational schema to GKB instance of the World.

```
CREATE TABLE GF_Pol_Name (
  gfn_id INT UNSIGNED NOT NULL PRIMARY KEY,
  gfn_name VARCHAR(255) NOT NULL,
  gfn_ascii VARCHAR(255) NOT NULL,
  gfn_cap VARCHAR(255) NOT NULL,
  gfn_lang char(2) NOT NULL,
  preferred_name char(1) NULL,
  official_name char(1) NULL,
  adjective char(1) NULL
);

CREATE TABLE GF_Pol_Type (
  gft_id char(10) PRIMARY KEY,
  gft_name VARCHAR(255) NOT NULL,
  gft_desc VARCHAR(255)
);

CREATE TABLE GF_Pol_Feature (
  gff_id INT UNSIGNED NOT NULL,
  gft_id char(10) NOT NULL REFERENCES GF_Type (gft_id) ON DELETE CASCADE,
  gfn_id INT UNSIGNED NOT NULL REFERENCES GF_Name (gfn_id) ON DELETE CASCADE,
  is_id char(4) NOT NULL REFERENCES Info_Source (is_id) ON DELETE CASCADE,
  PRIMARY KEY(gff_id,gft_id,gfn_id)
);

CREATE TABLE GF_Pol_Relationship_Type (
  gfrt_id CHAR(3) NOT NULL PRIMARY KEY,
  gfrt_name VARCHAR(255) NOT NULL,
  gfrt_desc VARCHAR(255)
);

CREATE TABLE GF_Pol_Relationship (
  gfr_id INT UNSIGNED NOT NULL,
  gff_id1 INT UNSIGNED NOT NULL REFERENCES GF_Feature (gff_id) ON DELETE CASCADE,
  gff_id2 INT UNSIGNED NOT NULL REFERENCES GF_Feature (gff_id) ON DELETE CASCADE,
  gfrt_id CHAR(3) NOT NULL REFERENCES GF_Relationship_Type(gfrt_id) ON DELETE CASCADE,
  is_id char(4) NOT NULL REFERENCES Info_Source (is_id) ON DELETE CASCADE,
  PRIMARY KEY(gff_id1,gff_id2)
);

CREATE TABLE GF_Pol_ISO_3166 (
  gf_iso INT UNSIGNED NOT NULL,
  gff_id INT UNSIGNED NOT NULL REFERENCES GF_Feature (gff_id) ON DELETE CASCADE,
  cod_iso VARCHAR(6) NOT NULL,
  pop INT NOT NULL
);

CREATE TABLE GF_Pol_Populated (
  gffp_id INT UNSIGNED NOT NULL,
  gff_id INT UNSIGNED NOT NULL REFERENCES GF_Feature (gff_id) ON DELETE CASCADE,
  pop INT NOT NULL
);

CREATE TABLE GF_Phy_Name (
  gfpn_id INT UNSIGNED NOT NULL PRIMARY KEY,
  gfpn_name VARCHAR(255) NOT NULL,
```

```
  gfpn_ascii VARCHAR(255) NOT NULL,
  gfpn_cap VARCHAR(255) NOT NULL,
  gfpn_lang char(2) NOT NULL,
  preferred_name char(1) NULL
);

CREATE TABLE GF_Phy_Type (
  gfpt_id char(10) PRIMARY KEY,
  gfpt_name VARCHAR(255) NOT NULL,
  gfpt_desc VARCHAR(255)
);

CREATE TABLE GF_Phy_Feature (
  gfpf_id INT UNSIGNED NOT NULL,
  gfpt_id char(10) NOT NULL REFERENCES GF_Type (gfpt_id) ON DELETE CASCADE,
  gfpn_id INT UNSIGNED NOT NULL REFERENCES GF_Name (gfpn_id) ON DELETE CASCADE,
  is_id char(4) NOT NULL REFERENCES Info_Source (is_id) ON DELETE CASCADE,
  PRIMARY KEY(gfpf_id,gfpt_id,gfpn_id)
);

CREATE TABLE GF_Phy_Relationship_Type (
  gfprt_id CHAR(3) NOT NULL PRIMARY KEY,
  gfprt_name VARCHAR(255) NOT NULL,
  gfprt_desc VARCHAR(255)
);

CREATE TABLE GF_Phy_Relationship (
  gfpr_id INT UNSIGNED NOT NULL,
  gfpf_id1 INT UNSIGNED NOT NULL REFERENCES GF_Feature (gfpf_id) ON DELETE CASCADE,
  gfpf_id2 INT UNSIGNED NOT NULL REFERENCES GF_Feature (gfpf_id) ON DELETE CASCADE,
  gfprt_id CHAR(3) NOT NULL REFERENCES GF_Relationship_Type(gfprt_id) ON DELETE CASCADE,
  is_id char(4) NOT NULL REFERENCES Info_Source (is_id) ON DELETE CASCADE,
  PRIMARY KEY(gfpf_id1,gfpf_id2)
);


CREATE TABLE ID_Relationship_Type (
  idrt_id CHAR(3) NOT NULL PRIMARY KEY,
  idrt_name VARCHAR(255) NOT NULL,
  idrt_desc VARCHAR(255)
);

CREATE TABLE ID_Relationship (
  idr_id INT UNSIGNED NOT NULL,
  id_id1 INT UNSIGNED NOT NULL,
  id_id2 INT UNSIGNED NOT NULL,
  type_rel CHAR(3) NOT NULL,
  is_id1 INT(3) NOT NULL REFERENCES Info_Source (is_id) ON DELETE CASCADE,
  is_id2 INT(3) NOT NULL REFERENCES Info_Source (is_id) ON DELETE CASCADE,
  PRIMARY KEY(id_id1,id_id2,is_id1,is_id2)
);
```

Appendix C - Vocabulary used in the ontology of Portugal

This appendix presents the vocabulary used in the ontology of Portugal. We define the classes and proprieties which are used in the file of the instances (data). Namespaces are also defined here.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!--  Declaração dos espaços de nomes -->
<rdf:RDF
  xmlns     = "http://xldb.di.fc.ul.pt/geo_net_pt01.owl#"
  xml:base  = "http://xldb.di.fc.ul.pt/geo_net_pt01.owl#"
  xmlns:gn = "http://xldb.di.fc.ul.pt/geo_net_pt01.owl#"
  xmlns:owl = "http://www.w3.org/2002/07/owl#"
  xmlns:rdf = "http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:rdfs = "http://www.w3.org/2000/01/rdf-schema#"
>
<!--  Declaração dos cabeçalhos -->
<owl:Ontology rdf:about="">
   <rdfs:comment>Descrição do vocabulário utilizado na ontologia geográfica de Portugal
   </rdfs:comment>
   <owl:priorVersion rdf:resource=""/>
   <rdfs:label>Ontologia Geográfica de Portugal</rdfs:label>
</owl:Ontology>

<!--  Definições das classes -->
<owl:Class rdf:ID="Geo_Feature">
  <rdfs:label>Feature geográfica - um objeto físico no domínio geográfico de Portugal
  </rdfs:label>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#geo_id"/>
      <owl:cardinality rdf:datatype="http://www.w3.org/2001/XMLSchema#nonNegativeInteger">1
      </owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#geo_name"/>
      <owl:cardinality rdf:datatype="http://www.w3.org/2001/XMLSchema#nonNegativeInteger">1
      </owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#alternative_name"/>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
       <owl:onProperty rdf:resource="#geo_type_id"/>
       <owl:allValuesFrom rdf:resource="#Geo_Type"/>
       <owl:cardinality rdf:datatype="http://www.w3.org/2001/XMLSchema#nonNegativeInteger">1
       </owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
       <owl:onProperty rdf:resource="#info_source_id"/>
         <owl:allValuesFrom>
          <owl:Class>
            <owl:oneOf rdf:parseType="Collection">
          <gn:Info_Source rdf:about="#INE"/>
          <gn:Info_Source rdf:about="#CTT"/>
```

```
          <gn:Info_Source rdf:about="#GAZ"/>
          <gn:Info_Source rdf:about="#WIKI"/>
          <gn:Info_Source rdf:about="#ANMP"/>
            </owl:oneOf>
          </owl:Class>
        </owl:allValuesFrom>
        <owl:cardinality rdf:datatype="http://www.w3.org/2001/XMLSchema#nonNegativeInteger">1
        </owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#related_to"/>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#population"/>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#latitude"/>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#longitude"/>
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>

<owl:Class rdf:ID="Geo_Name">
  <rdfs:label>Nomes utilizados no domínio administrativo geográfico de Portugal</rdfs:label>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#geo_name"/>
      <owl:cardinality rdf:datatype="http://www.w3.org/2001/XMLSchema#nonNegativeInteger">1
      </owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
       <owl:onProperty rdf:resource="#info_source_id"/>
         <owl:allValuesFrom>
          <owl:Class>
            <owl:oneOf rdf:parseType="Collection">
          <gn:Info_Source rdf:about="#INE"/>
          <gn:Info_Source rdf:about="#CTT"/>
          <gn:Info_Source rdf:about="#GAZ"/>
          <gn:Info_Source rdf:about="#WIKI"/>
          <gn:Info_Source rdf:about="#ANMP"/>
            </owl:oneOf>
          </owl:Class>
        </owl:allValuesFrom>
        <owl:cardinality rdf:datatype="http://www.w3.org/2001/XMLSchema#nonNegativeInteger">1
        </owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>

<owl:Class rdf:ID="Info_Source">
```

```
    <rdfs:label>Fontes de informação de onde provêm os dados da ontologia geográfica de Portugal
    </rdfs:label>
    <owl:oneOf rdf:parseType="Collection">
      <owl:Thing rdf:about="#INE"/>
      <owl:Thing rdf:about="#CTT"/>
      <owl:Thing rdf:about="#GAZ"/>
      <owl:Thing rdf:about="#WIKI"/>
      <owl:Thing rdf:about="#ANMP"/>
      <owl:Thing rdf:about="#FCCN"/>
      <owl:Thing rdf:about="#PT4"/>
      <owl:Thing rdf:about="#PT5"/>
    </owl:oneOf>
</owl:Class>

<owl:Class rdf:ID="Geo_Relationship">
    <rdfs:label>Relacionamentos semânticos entre as features geográficas</rdfs:label>
    <rdfs:subClassOf>
        <owl:Restriction>
        <owl:onProperty rdf:resource="#rel_type_id"/>
        <owl:allValuesFrom>
            <owl:Class>
                <owl:oneOf rdf:parseType="Collection">
                    <owl:Thing rdf:about="#PRT"/>
                    <owl:Thing rdf:about="#ADJ"/>
                    <owl:Thing rdf:about="#SBP"/>
                </owl:oneOf>
            </owl:Class>
        </owl:allValuesFrom>
        </owl:Restriction>
    </rdfs:subClassOf>
    <rdfs:subClassOf>
      <owl:Restriction>
         <owl:onProperty rdf:resource="#geo_id"/>
      </owl:Restriction>
    </rdfs:subClassOf>
</owl:Class>

<owl:Class rdf:ID="Geo_Type">
  <rdfs:label>Tipos de features utilizadas na ontologia geográfica de Portugal</rdfs:label>
  <owl:oneOf rdf:parseType="Collection">
    <owl:Thing rdf:about="#ACE"/>
    <owl:Thing rdf:about="#ADR"/>
    <owl:Thing rdf:about="#ALA"/>
    <owl:Thing rdf:about="#AVE"/>
    <owl:Thing rdf:about="#AZI"/>
    <owl:Thing rdf:about="#BAI"/>
    <owl:Thing rdf:about="#BEC"/>
    <owl:Thing rdf:about="#CAI"/>
    <owl:Thing rdf:about="#CAL"/>
    <owl:Thing rdf:about="#CAM"/>
    <owl:Thing rdf:about="#CAN"/>
    <owl:Thing rdf:about="#CAR"/>
    <owl:Thing rdf:about="#CDP"/>
    <owl:Thing rdf:about="#CON"/>
    <owl:Thing rdf:about="#CPO"/>
    <owl:Thing rdf:about="#CTO"/>
    <owl:Thing rdf:about="#DST"/>
    <owl:Thing rdf:about="#ENH"/>
    <owl:Thing rdf:about="#ESC"/>
    <owl:Thing rdf:about="#EST"/>
    <owl:Thing rdf:about="#FRG"/>
    <owl:Thing rdf:about="#ILH"/>
```

```
    <owl:Thing rdf:about="#JAR"/>
    <owl:Thing rdf:about="#LAD"/>
    <owl:Thing rdf:about="#LOC"/>
    <owl:Thing rdf:about="#LOT"/>
    <owl:Thing rdf:about="#LRG"/>
    <owl:Thing rdf:about="#LUG"/>
    <owl:Thing rdf:about="#MON"/>
    <owl:Thing rdf:about="#NT1"/>
    <owl:Thing rdf:about="#NT2"/>
    <owl:Thing rdf:about="#NT3"/>
    <owl:Thing rdf:about="#OUT"/>
    <owl:Thing rdf:about="#PAI"/>
    <owl:Thing rdf:about="#PAR"/>
    <owl:Thing rdf:about="#PAS"/>
    <owl:Thing rdf:about="#PAT"/>
    <owl:Thing rdf:about="#PRA"/>
    <owl:Thing rdf:about="#PTA"/>
    <owl:Thing rdf:about="#PTE"/>
    <owl:Thing rdf:about="#QUE"/>
    <owl:Thing rdf:about="#QUI"/>
    <owl:Thing rdf:about="#REC"/>
    <owl:Thing rdf:about="#RLA"/>
    <owl:Thing rdf:about="#ROT"/>
    <owl:Thing rdf:about="#RPA"/>
    <owl:Thing rdf:about="#RUA"/>
    <owl:Thing rdf:about="#SIT"/>
    <owl:Thing rdf:about="#TER"/>
    <owl:Thing rdf:about="#TRV"/>
    <owl:Thing rdf:about="#URB"/>
    <owl:Thing rdf:about="#VER"/>
    <owl:Thing rdf:about="#VIA"/>
    <owl:Thing rdf:about="#VIE"/>
    <owl:Thing rdf:about="#VLE"/>
    <owl:Thing rdf:about="#ZNA"/>
    <owl:Thing rdf:about="#ZON"/>
  </owl:oneOf>
</owl:Class>

<!--  DatatypeProperty Definitions -->
<owl:DatatypeProperty rdf:ID="geo_id">
    <rdfs:label>Código identificador de uma feature geográfica</rdfs:label>
    <rdfs:domain rdf:resource="#Geo_Feature"/>
    <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#nonNegativeInteger"/>
</owl:DatatypeProperty>

<owl:DatatypeProperty rdf:ID="population">
    <rdfs:label>Número de pessoas que residem em uma feature geográfica. Nessa ontologia
    esse número é atribuído as features do tipo concelho</rdfs:label>
    <rdfs:domain rdf:resource="#Geo_Feature"/>
    <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#nonNegativeInteger"/>
</owl:DatatypeProperty>

<owl:DatatypeProperty rdf:ID="latitude">
    <rdfs:domain rdf:resource="#Geo_Feature"/>
    <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#double"/>
</owl:DatatypeProperty>

<owl:DatatypeProperty rdf:ID="longitude">
    <rdfs:domain rdf:resource="#Geo_Feature"/>
    <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#double"/>
</owl:DatatypeProperty>
```

```
<owl:DatatypeProperty rdf:ID="geo_name">
    <rdfs:label>Propriedade utilizada para associar um nome a cada feature geográfica
    </rdfs:label>
    <rdfs:domain rdf:resource="#Geo_Feature"/>
    <rdfs:domain rdf:resource="#Geo_Name"/>
    <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#string"/>
</owl:DatatypeProperty>

<owl:DatatypeProperty rdf:ID="rel_type_id">
    <rdfs:label>Código identificador dos relacionamentos semânticos entre as features
    geográficas</rdfs:label>
    <rdfs:domain rdf:resource="#Geo_Relationship"/>
    <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#string"/>
</owl:DatatypeProperty>

<!--  ObjectProperty Definitions -->
<owl:ObjectProperty rdf:ID="alternative_name">
    <rdfs:domain rdf:resource="#Geo_Feature"/>
    <rdfs:range rdf:resource="#Geo_Name"/>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="info_source_id">
    <rdfs:label>Código identificador da fonte de informação geográfica</rdfs:label>
    <rdfs:domain rdf:resource="#Geo_Feature"/>
    <rdfs:domain rdf:resource="#Geo_Name"/>
    <rdfs:range rdf:resource="#Info_Source"/>
</owl:ObjectProperty>

<!-- Begin of the classes declarations from Net domain -->
<owl:Class rdf:ID="Net_Feature">
  <rdfs:label>Feature internet - composta ao menos por um nome, um tipo e uma fonte de
  informação</rdfs:label>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#net_id"/>
      <owl:cardinality rdf:datatype="http://www.w3.org/2001/XMLSchema#nonNegativeInteger">1
      </owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#net_name"/>
      <owl:cardinality rdf:datatype="http://www.w3.org/2001/XMLSchema#nonNegativeInteger">1
      </owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#net_type_id"/>
      <owl:allValuesFrom rdf:resource="#Net_Type"/>
      <owl:cardinality rdf:datatype="http://www.w3.org/2001/XMLSchema#nonNegativeInteger">1
      </owl:cardinality>
     </owl:Restriction>
   </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
       <owl:onProperty rdf:resource="#info_source_id"/>
         <owl:allValuesFrom>
          <owl:Class>
            <owl:oneOf rdf:parseType="Collection">
              <gn:Info_Source rdf:about="#PT4"/>
              <gn:Info_Source rdf:about="#PT5"/>
```

```
            <gn:Info_Source rdf:about="#FCCN"/>
          </owl:oneOf>
        </owl:Class>
      </owl:allValuesFrom>
    <owl:cardinality rdf:datatype="http://www.w3.org/2001/XMLSchema#nonNegativeInteger">1
    </owl:cardinality>
  </owl:Restriction>
 </rdfs:subClassOf>
 <rdfs:subClassOf>
  <owl:Restriction>
    <owl:onProperty rdf:resource="#has_scope"/>
  </owl:Restriction>
 </rdfs:subClassOf>
</owl:Class>

<owl:Class rdf:ID="Net_Type">
 <rdfs:label>Tipo de feature internet: um domínio, um site ou uma página</rdfs:label>
 <owl:oneOf rdf:parseType="Collection">
  <owl:Thing rdf:about="#DOM"/>
  <owl:Thing rdf:about="#STE"/>
 </owl:oneOf>
</owl:Class>

<!--  DatatypeProperty Definitions -->
<owl:DatatypeProperty rdf:ID="net_id">
 <rdfs:label>código identificador da feture internet</rdfs:label>
 <rdfs:domain rdf:resource="#Net_Feature"/>
 <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#nonNegativeInteger"/>
</owl:DatatypeProperty>

<owl:DatatypeProperty rdf:ID="net_name">
 <rdfs:label>Nome da feature internet</rdfs:label>
 <rdfs:domain rdf:resource="#Net_Feature"/>
 <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#string"/>
</owl:DatatypeProperty>

<owl:DatatypeProperty rdf:ID="ip_number">
 <rdfs:label>Número IP (Internet Protocol) de um dominio ou um site</rdfs:label>
 <rdfs:domain rdf:resource="#Net_Feature"/>
 <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#float"/>
</owl:DatatypeProperty>

<!--  ObjectProperty Definitions -->
<!--  The owl;FunctionalProperty asserts that the has_scope property has at most one
Geo_Feature -->
<owl:ObjectProperty rdf:ID="has_scope">
 <rdfs:label>Propriedade utilizada para associar um código de uma feature geográfica atribuindo
 um âmbito para uma feature internet do tipo site</rdfs:label>
 <rdf:type rdf:resource="owl;FunctionalProperty"/>
 <rdfs:domain rdf:resource="#Net_Feature"/>
 <rdfs:range rdf:resource="#Geo_Feature"/>
</owl:ObjectProperty>

<Info_Source rdf:ID="INE">
 <rdfs:label>Instituto Nacional de Estatística</rdfs:label>
</Info_Source>

<Info_Source rdf:ID="CTT">
 <rdfs:label>Correios Telégrafos e Telefones</rdfs:label>
</Info_Source>

<Info_Source rdf:ID="GAZ">
```

```
  <rdfs:label>Gazetteer-www.calle.com</rdfs:label>
</Info_Source>

<Info_Source rdf:ID="WIKI">
  <rdfs:label>Wikipedia</rdfs:label>
</Info_Source>

<Info_Source rdf:ID="ANMP">
  <rdfs:label>Associação Nacional de Municípios Portugueses</rdfs:label>
</Info_Source>

<Info_Source rdf:ID="FCCN">
  <rdfs:label>Fundação para a Computação Científica Nacional</rdfs:label>
</Info_Source>

<Info_Source rdf:ID="PT4">
  <rdfs:label>Web sites found by tumba!</rdfs:label>
</Info_Source>

<Info_Source rdf:ID="PT5">
  <rdfs:label>Web sites found by tumba!</rdfs:label>
</Info_Source>

<Geo_Type rdf:ID="ACE">
  <rdfs:label>adm:arruamento:acesso</rdfs:label>
</Geo_Type>

<Geo_Type rdf:ID="ADR">
  <rdfs:label>adm:arruamento:adro</rdfs:label>
</Geo_Type>

<Geo_Type rdf:ID="ALA">
  <rdfs:label>adm:arruamento:alameda</rdfs:label>
</Geo_Type>

<Geo_Type rdf:ID="AVE">
  <rdfs:label>adm:arruamento:avenida</rdfs:label>
</Geo_Type>

<Geo_Type rdf:ID="AZI">
  <rdfs:label>adm:arruamento:azinhaga</rdfs:label>
</Geo_Type>

<Geo_Type rdf:ID="BAI">
  <rdfs:label>adm:arruamento:bairro</rdfs:label>
</Geo_Type>

<Geo_Type rdf:ID="BEC">
  <rdfs:label>adm:arruamento:beco</rdfs:label>
</Geo_Type>

<Geo_Type rdf:ID="CAI">
  <rdfs:label>adm:arruamento:cais</rdfs:label>
</Geo_Type>

<Geo_Type rdf:ID="CAL">
  <rdfs:label>adm:arruamento:calçada</rdfs:label>
</Geo_Type>

<Geo_Type rdf:ID="CAM">
  <rdfs:label>adm:arruamento:caminho</rdfs:label>
</Geo_Type>
```

```
<Geo_Type rdf:ID="CAN">
  <rdfs:label>adm:arruamento:canada</rdfs:label>
</Geo_Type>

<Geo_Type rdf:ID="CAR">
  <rdfs:label>adm:arruamento:carreira</rdfs:label>
</Geo_Type>

<Geo_Type rdf:ID="CDP">
  <rdfs:label>adm:codigo_postal</rdfs:label>
</Geo_Type>

<Geo_Type rdf:ID="CON">
  <rdfs:label>adm:concelho</rdfs:label>
</Geo_Type>

<Geo_Type rdf:ID="CPO">
  <rdfs:label>adm:arruamento:campo</rdfs:label>
</Geo_Type>

<Geo_Type rdf:ID="CTO">
  <rdfs:label>adm:arruamento:canto</rdfs:label>
</Geo_Type>

<Geo_Type rdf:ID="DST">
  <rdfs:label>adm:distrito</rdfs:label>
</Geo_Type>

<Geo_Type rdf:ID="ENH">
  <rdfs:label>adm:arruamento:escadinhas</rdfs:label>
</Geo_Type>

<Geo_Type rdf:ID="ESC">
  <rdfs:label>adm:arruamento:escadas</rdfs:label>
</Geo_Type>

<Geo_Type rdf:ID="EST">
  <rdfs:label>adm:arruamento:estrada</rdfs:label>
</Geo_Type>

<Geo_Type rdf:ID="FRG">
  <rdfs:label>adm:freguesia</rdfs:label>
</Geo_Type>

<Geo_Type rdf:ID="ILH">
  <rdfs:label>adm:ilha</rdfs:label>
</Geo_Type>

<Geo_Type rdf:ID="JAR">
  <rdfs:label>adm:arruamento:jardim</rdfs:label>
</Geo_Type>

<Geo_Type rdf:ID="LAD">
  <rdfs:label>adm:arruamento:ladeira</rdfs:label>
</Geo_Type>

<Geo_Type rdf:ID="LOC">
  <rdfs:label>adm:localidade</rdfs:label>
</Geo_Type>

<Geo_Type rdf:ID="LOT">
```

```
  <rdfs:label>adm:arruamento:loteamento</rdfs:label>
</Geo_Type>

<Geo_Type rdf:ID="LRG">
  <rdfs:label>adm:arruamento:largo</rdfs:label>
</Geo_Type>

<Geo_Type rdf:ID="LUG">
  <rdfs:label>adm:arruamento:lugar</rdfs:label>
</Geo_Type>

<Geo_Type rdf:ID="MON">
  <rdfs:label>adm:arruamento:monte</rdfs:label>
</Geo_Type>

<Geo_Type rdf:ID="NT1">
  <rdfs:label>adm:NUT1</rdfs:label>
</Geo_Type>

<Geo_Type rdf:ID="NT2">
  <rdfs:label>adm:NUT2</rdfs:label>
</Geo_Type>

<Geo_Type rdf:ID="NT3">
  <rdfs:label>adm:NUT3</rdfs:label>
</Geo_Type>

<Geo_Type rdf:ID="OUT">
  <rdfs:label>adm:arruamento:outro</rdfs:label>
</Geo_Type>

<Geo_Type rdf:ID="PAI">
  <rdfs:label>adm:pais</rdfs:label>
</Geo_Type>

<Geo_Type rdf:ID="PAR">
  <rdfs:label>adm:arruamento:parque</rdfs:label>
</Geo_Type>

<Geo_Type rdf:ID="PAS">
  <rdfs:label>adm:arruamento:passeio</rdfs:label>
</Geo_Type>

<Geo_Type rdf:ID="PAT">
  <rdfs:label>adm:arruamento:pátio</rdfs:label>
</Geo_Type>

<Geo_Type rdf:ID="PRA">
  <rdfs:label>adm:arruamento:praça</rdfs:label>
</Geo_Type>

<Geo_Type rdf:ID="PTA">
  <rdfs:label>adm:arruamento:praceta</rdfs:label>
</Geo_Type>

<Geo_Type rdf:ID="PTE">
  <rdfs:label>adm:arruamento:ponte</rdfs:label>
</Geo_Type>

<Geo_Type rdf:ID="QUE">
  <rdfs:label>adm:arruamento:quelha</rdfs:label>
</Geo_Type>
```

```
<Geo_Type rdf:ID="QUI">
  <rdfs:label>adm:arruamento:quinta</rdfs:label>
</Geo_Type>

<Geo_Type rdf:ID="REC">
  <rdfs:label>adm:arruamento:recanto</rdfs:label>
</Geo_Type>

<Geo_Type rdf:ID="RLA">
  <rdfs:label>adm:arruamento:ruela</rdfs:label>
</Geo_Type>

<Geo_Type rdf:ID="ROT">
  <rdfs:label>adm:arruamento:rotunda</rdfs:label>
</Geo_Type>

<Geo_Type rdf:ID="RPA">
  <rdfs:label>adm:arruamento:rampa</rdfs:label>
</Geo_Type>

<Geo_Type rdf:ID="RUA">
  <rdfs:label>adm:arruamento:rua</rdfs:label>
</Geo_Type>

<Geo_Type rdf:ID="SIT">
  <rdfs:label>adm:arruamento:sítio</rdfs:label>
</Geo_Type>

<Geo_Type rdf:ID="TER">
  <rdfs:label>adm:arruamento:terreiro</rdfs:label>
</Geo_Type>

<Geo_Type rdf:ID="TRV">
  <rdfs:label>adm:arruamento:travessa</rdfs:label>
</Geo_Type>

<Geo_Type rdf:ID="URB">
  <rdfs:label>adm:arruamento:urbanização</rdfs:label>
</Geo_Type>

<Geo_Type rdf:ID="VER">
  <rdfs:label>adm:arruamento:vereda</rdfs:label>
</Geo_Type>

<Geo_Type rdf:ID="VIA">
  <rdfs:label>adm:arruamento:via</rdfs:label>
</Geo_Type>

<Geo_Type rdf:ID="VIE">
  <rdfs:label>adm:arruamento:viela</rdfs:label>
</Geo_Type>

<Geo_Type rdf:ID="VLE">
  <rdfs:label>adm:arruamento:vale</rdfs:label>
</Geo_Type>

<Geo_Type rdf:ID="ZNA">
  <rdfs:label>adm:arruamento:zona</rdfs:label>
</Geo_Type>

<Geo_Type rdf:ID="ZON">
```

```
  <rdfs:label>adm:zona</rdfs:label>
</Geo_Type>

<Geo_Relationship rdf:ID="PRT">
  <rdfs:label>parteDe</rdfs:label>
</Geo_Relationship>

<Geo_Relationship rdf:ID="ADJ">
  <rdfs:label>adjacente</rdfs:label>
</Geo_Relationship>

<Geo_Relationship rdf:ID="SBP">
  <rdfs:label>sobreposição</rdfs:label>
</Geo_Relationship>

<Net_Type rdf:ID="DOM">
  <rdfs:label>domínio web</rdfs:label>
</Net_Type>

<Net_Type rdf:ID="STE">
  <rdfs:label>sítio web</rdfs:label>
</Net_Type>

</rdf:RDF>
```

Appendix D - Vocabulary used in the geographic ontology of the World

This appendix presents the vocabulary used in the ontology of the World. We define the classes and proprieties which are used in the file of the instances (data). Namespaces are also defined here.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!--  Declaration of Namespaces -->
<rdf:RDF
  xmlns     = "http://xldb.di.fc.ul.pt/geo_world.owl#"
  xml:base  = "http://xldb.di.fc.ul.pt/geo_world.owl#"
  xmlns:gw = "http://xldb.di.fc.ul.pt/geo_world.owl#"
  xmlns:owl = "http://www.w3.org/2002/07/owl#"
  xmlns:rdf = "http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:rdfs = "http://www.w3.org/2000/01/rdf-schema#"
>
<!--  Declaration of Headers -->
<owl:Ontology rdf:about="">
   <rdfs:comment>Description of classes and properties of Geographic World Ontology
   </rdfs:comment>
   <owl:priorVersion rdf:resource=""/>
   <rdfs:label>Geographic World Ontology</rdfs:label>
</owl:Ontology>

<!--  Classes Definitions -->
<!--  The owl:cardinality restriction asserts that the property been defined has
exactly one value -->
<owl:Class rdf:ID="Geo_Feature">
  <rdfs:label>Geographic feature</rdfs:label>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#geo_id"/>
      <owl:cardinality rdf:datatype="http://www.w3.org/2001/XMLSchema#nonNegativeInteger">1
      </owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#geo_name"/>
      <owl:cardinality rdf:datatype="http://www.w3.org/2001/XMLSchema#nonNegativeInteger">1
      </owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#common_name"/>
      <owl:cardinality rdf:datatype="http://www.w3.org/2001/XMLSchema#nonNegativeInteger">1
      </owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#official_name"/>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
       <owl:onProperty rdf:resource="#geo_type_id"/>
       <owl:allValuesFrom rdf:resource="#Geo_Type"/>
       <owl:cardinality rdf:datatype="http://www.w3.org/2001/XMLSchema#nonNegativeInteger">1
       </owl:cardinality>
```

```
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
       <owl:onProperty rdf:resource="#info_source_id"/>
         <owl:allValuesFrom>
          <owl:Class>
            <owl:oneOf rdf:parseType="Collection">
          <gn:Info_Source rdf:about="#GAZ"/>
          <gn:Info_Source rdf:about="#WIKI"/>
            </owl:oneOf>
          </owl:Class>
         </owl:allValuesFrom>
         <owl:cardinality rdf:datatype="http://www.w3.org/2001/XMLSchema#nonNegativeInteger">1
         </owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#related_to"/>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#population"/>
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>

<owl:Class rdf:ID="Geo_Name">
  <rdfs:label>Geographic names</rdfs:label>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#geo_name"/>
      <owl:cardinality rdf:datatype="http://www.w3.org/2001/XMLSchema#nonNegativeInteger">1
      </owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
       <owl:onProperty rdf:resource="#info_source_id"/>
         <owl:allValuesFrom>
          <owl:Class>
            <owl:oneOf rdf:parseType="Collection">
          <gn:Info_Source rdf:about="#WIKI"/>
          <gn:Info_Source rdf:about="#WGAZ"/>
            </owl:oneOf>
          </owl:Class>
         </owl:allValuesFrom>
         <owl:cardinality rdf:datatype="http://www.w3.org/2001/XMLSchema#nonNegativeInteger">1
         </owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#lang"/>
      <owl:cardinality rdf:datatype="http://www.w3.org/2001/XMLSchema#nonNegativeInteger">1
      </owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>
```

```
<owl:Class rdf:ID="Info_Source">
    <rdfs:label>Information Sources</rdfs:label>
    <owl:oneOf rdf:parseType="Collection">
      <owl:Thing rdf:about="#WIKI"/>
      <owl:Thing rdf:about="#WGAZ"/>
    </owl:oneOf>
</owl:Class>

<owl:Class rdf:ID="Geo_Relationship">
    <rdfs:label>Relationships among features</rdfs:label>
    <rdfs:subClassOf>
        <owl:Restriction>
        <owl:onProperty rdf:resource="#rel_type_id"/>
        <owl:allValuesFrom>
            <owl:Class>
                <owl:oneOf rdf:parseType="Collection">
                    <owl:Thing rdf:about="#PRT"/>
                    <owl:Thing rdf:about="#ADJ"/>
                </owl:oneOf>
            </owl:Class>
        </owl:allValuesFrom>
        </owl:Restriction>
    </rdfs:subClassOf>
    <rdfs:subClassOf>
      <owl:Restriction>
         <owl:onProperty rdf:resource="#geo_id"/>
      </owl:Restriction>
    </rdfs:subClassOf>
</owl:Class>

<owl:Class rdf:ID="Geo_Type">
  <rdfs:label>Geographic features types</rdfs:label>
  <owl:oneOf rdf:parseType="Collection">
    <owl:Thing rdf:about="#ISO-3166-1"/>
    <owl:Thing rdf:about="#ISO-3166-2"/>
    <owl:Thing rdf:about="#CITY-CAP"/>
    <owl:Thing rdf:about="#PLACE"/>
    <owl:Thing rdf:about="#ADM_DIV"/>
    <owl:Thing rdf:about="#AGGLO"/>
    <owl:Thing rdf:about="#REG"/>
    <owl:Thing rdf:about="#PLAN"/>
    <owl:Thing rdf:about="#CONT"/>
    <owl:Thing rdf:about="#SEA"/>
    <owl:Thing rdf:about="#LAKE"/>
  </owl:oneOf>
</owl:Class>

<!--  DatatypeProperty Definitions -->
<owl:DatatypeProperty rdf:ID="geo_id">
    <rdfs:label>Geographic feature identifier</rdfs:label>
    <rdfs:domain rdf:resource="#Geo_Feature"/>
    <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#nonNegativeInteger"/>
</owl:DatatypeProperty>

<owl:DatatypeProperty rdf:ID="population">
    <rdfs:label>Number of people in the geographic feature</rdfs:label>
    <rdfs:domain rdf:resource="#Geo_Feature"/>
    <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#nonNegativeInteger"/>
</owl:DatatypeProperty>

<owl:DatatypeProperty rdf:ID="geo_name">
    <rdfs:label>Geographic name</rdfs:label>
```

```
    <rdfs:domain rdf:resource="#Geo_Name"/>
    <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#string"/>
</owl:DatatypeProperty>

<owl:DatatypeProperty rdf:ID="lang">
    <rdfs:label>Language (idiom) in which the name is written</rdfs:label>
    <rdfs:domain rdf:resource="#Geo_Name"/>
    <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#string"/>
</owl:DatatypeProperty>

<owl:DatatypeProperty rdf:ID="rel_type_id">
    <rdfs:label>Identifier of the geographic relationship type</rdfs:label>
    <rdfs:domain rdf:resource="#Geo_Relationship"/>
    <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#string"/>
</owl:DatatypeProperty>

<!--  ObjectProperty Definitions -->
<owl:ObjectProperty rdf:ID="common_name">
    <rdfs:domain rdf:resource="#Geo_Feature"/>
    <rdfs:range rdf:resource="#Geo_Name"/>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="official_name">
    <rdfs:domain rdf:resource="#Geo_Feature"/>
    <rdfs:range rdf:resource="#Geo_Name"/>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="info_source_id">
    <rdfs:label>Information source identifier</rdfs:label>
    <rdfs:domain rdf:resource="#Geo_Feature"/>
    <rdfs:domain rdf:resource="#Geo_Name"/>
    <rdfs:range rdf:resource="#Info_Source"/>
</owl:ObjectProperty>

<Info_Source rdf:ID="WIKI">
  <rdfs:label>Wikipedia</rdfs:label>
</Info_Source>

<Info_Source rdf:ID="WGAZ">
  <rdfs:label>World Gazetteer</rdfs:label>
</Info_Source>

<Geo_Type rdf:ID="ISO-3166-1">
  <rdfs:label>ISO-3166-1</rdfs:label>
</Geo_Type>

<Geo_Type rdf:ID="ISO-3166-2">
  <rdfs:label>ISO-3166-2</rdfs:label>
</Geo_Type>

<Geo_Type rdf:ID="CITY-CAP">
  <rdfs:label>Name of the city</rdfs:label>
</Geo_Type>

<Geo_Type rdf:ID="PLACE">
  <rdfs:label>Name of a geographic place</rdfs:label>
</Geo_Type>

<Geo_Type rdf:ID="ADM_DIV">
  <rdfs:label>Administrative division</rdfs:label>
</Geo_Type>
```

```
<Geo_Type rdf:ID="AGGLO">
  <rdfs:label>Agglomeration</rdfs:label>
</Geo_Type>

<Geo_Type rdf:ID="REG">
  <rdfs:label>Region</rdfs:label>
</Geo_Type>

<Geo_Type rdf:ID="PLAN">
  <rdfs:label>Planet</rdfs:label>
</Geo_Type>

<Geo_Type rdf:ID="CONT">
  <rdfs:label>Continent</rdfs:label>
</Geo_Type>

<Geo_Type rdf:ID="SEA"/>

<Geo_Type rdf:ID="LAKE"/>

</rdf:RDF>
```