

The Geo-Net-PT/Yahoo! GeoPlanet™ concordance

João D Ferreira, David S Batista, Francisco M Couto,
Mário J Silva

DI-FCUL-TR-2010-05

DOI:10455/6677

(<http://hdl.handle.net/10455/6677>)

October 2010



Published at Docs.DI (<http://docs.di.fc.ul.pt/>), the repository of the
Department of Informatics of the University of Lisbon, Faculty of Sciences.

The Geo-Net-PT/Yahoo! GeoPlanetTM concordance

João D Ferreira

joao.ferreira@lasige.di.fc.ul.pt

David S Batista

dsbatista@xldb.di.fc.ul.pt

Francisco M Couto

fjmc@di.fc.ul.pt

Mário J Silva

mjs@di.fc.ul.pt

October 14, 2010

Abstract

Background: Geo-Net-PT is a geospatial ontology representing the Portuguese territory and the relations between the several locations within it. Yahoo! GeoPlanetTM is a geospatial ontology that covers the whole world. To interlink the two ontologies and reduce the effects of repeated information, we propose an automatic alignment between their administrative parts, based on name similarity and physical closeness.

Results: After the matching process, we found 18,588 matches, corresponding to 35% of the considered features in Geo-Net-PT and to 71% in Yahoo! GeoPlanetTM. Among these, there are correct matches for each of the 18 districts and 308 municipalities. Only 1% of the matches failed the validation process.

Conclusions: This alignment represents a step further for the exploitation of geospatial ontologies, since it shows that it is possible to effectively cross-link geographical references between a broader ontology, such as Yahoo GeoPlanetTM, and a narrower yet more comprehensive one, such as Geo-Net-PT, providing the broader ontology with greater detail and complementary information.

1 Introduction

Geospatial ontologies are a representation of a geographical knowledge domain in a machine-readable form, containing geographical features and the relations between them, such as “Lisbon” *part-of* “Portugal” or “Lisbon district” *adjacent-to* “Setúbal district.”

A geospatial ontology can be useful for projects where geographical information plays an important role. One of the objectives of such ontologies is to provide a standard to annotate and connect different resources, like news or epidemiological models [5], or for automatic reasoning over the relationships between geographical locations, as advocated by the Semantic Web [1].

The geospatial information on Earth is so wide that it is beneficial to have ontologies spanning a range of specificity: some, wider in range, represent the whole Earth in a coherent structure that can be used for broader purposes, referring several countries or cities; smaller ontologies, with more specific spatial

information, are ideal for localized problems like local news. An example of a wide geospatial ontology is Yahoo! GeoPlanetTM [11] (YGP), an ontology that covers places from all around the world; Geo-Net-PT [6] (GNP) is an example of the second case, since it is a geospatial ontology spanning only the Portuguese territory [6].

The existence of more than one ontology, however, hinders the standardization process, since a user wishing to make a reference to a particular location will have to make a choice between the available ontologies, which will lead to cases where some annotations point to different references but were ultimately intended to refer to the same entity. One possible solution for this problem is the development of ontology alignments [3], which are sets of links between the terms in each ontology, such that by referring to one of them it is possible to reach the related terms in the other ontologies. Consider a document that has been annotated with the term “Lisboa” from Geo-Net-PT; knowing that this term and “Lisbon” from Yahoo! GeoPlanetTM refer to the same entity, removes the ambiguity. With this information, a search engine can actually retrieve that same document even if the query term is the term from Yahoo! GeoPlanetTM.

This work presents an alignment between GNP and the part of YGP covering the Portuguese territory. YGP is already enriched with alignments to other geographical services (not all of them ontologies), such as geonames.org, IATA reference (a reference of international airports) or Canada Post (list of Canadian postal codes). This alignment complements that information.

The last few years have seen an increase in the effort made to standardize and link together resources in the Web. LinkedData [7] is an umbrella project that aims at producing a recommendation for the exposure, distribution and connection of information on the Semantic Web. The recommended way of doing so is by linking together URIs through triples in an RDF (Resource Description Framework) file. In the same spirit, we have created an RDF [8] file containing all the matching pairs found in the two ontologies.

These triplets link GNP features to YGP features through the use of SKOS (Simple Knowledge Organization System) properties [10]. SKOS is a system that develops a standard way to represent knowledge using RDF. It provides a series of relations, some of which are appropriate for mappings.

Throughout the paper we make references to terms in both of these ontologies and the relations between them. We write terms in **sans-serif** and relations in a *slanted* typeface. Terms are usually accompanied with a qualifier, or qname: **gnp:** for GNP and **ygp:** for YGP. If the term is surrounded by double quotation marks, it refers to a location by name; if the term is a plain number, it refers to the identifier. Moreover, terms in geospatial ontologies are categorized into types: Countries, Municipalities, etc. Types are written in **bold**. A reference to a term in an ontology is done according to this format: **qname:ID** (“Name”, **type**).

The rest of this paper is structured as follows: Section 2 describes the two ontologies which will be aligned; in Section 3 we describe the methodology used to perform the alignment; Section 4 shows the results and statistics of the alignment and describes how the alignment is made available; finally, in Section 5 we draw the final conclusions.

2 Geographical Ontologies

Both GNP and YGP are organized as directed acyclic graphs based on the *part-of* relationships between the included geographic features [6, 11]. Each feature has a number of properties, of which two are mandatory: name and type. The name is usually localized into the language of the country; the type is based on the administrative regions of the country. For instance, the term Lisbon can refer to a district, a municipality or a city. Furthermore, other geographical properties may be present, like geospatial coordinates, bounding boxes, area or population, although these are not always available.

2.1 Geo-Net-PT

Geo-Net-PT is available from Linguateca at <http://linguateca/Geo-Net-PT>. This ontology is licensed under a Creative Commons Attribution 3.0 License (CC-BY). A copy of this license is available at <http://creativecommons.org/licenses/by/3.0/>.

GNP is a public geographic ontology covering the territory of Portugal. It is divided in two domains: administrative and physical. The administrative domain contains the administrative divisions of the territory and the physical domain includes physical geography features, such as natural regions and man-made spots [6]. Tables 1a and 1b present statistics organized by feature type.

Relationships of type *part-of* and *adjacent-to* exist between features in each domain and also between the two domains. Table 1c shows the number and type of intra- and inter-domain relationships. The feature types in GNP’s administrative domain are themselves arranged in a hierarchical manner through *part-of* and *adjacent-to* relationships, as showed in Figure 1.

2.2 GeoPlanetTM

Yahoo! GeoPlanetTM [11] is a world coverage geographic ontology. Each place in this ontology is identified by a unique identifier dubbed *Where On Earth Identifier* (WOEID). As of August 2010, each place in the Portuguese part of this ontology can have one of the types listed in Table 2. YGP makes a distinction between *official* administrative places and *informal* places, such as colloquial places and historical administrative places.

Of all the relationships between features in YGP, we only use the parent-child relationships, defined as “the direct inferiors to a given place.” This is equivalent to a *part-of* relation. In GeoPlanetTM, places have only one parent. There are 23,481 relations, one for every feature, connecting it with its parent, minus the absent relation of the root of the ontology, *ygp*: “Portugal”. Geographic feature types in YGP are also hierarchically related to one another through *part-of* relationships, as shown in Figure 2.

In the beginning of the work, we retrieve all information from YGP concerning the Portuguese territory. To do this, we query Yahoo!’s GeoPlanetTM web service in order to retrieve *ygp*:23424925 “Portugal”, its children and the children of all its descendants, recursively, until no more features are found.

Table 1: Characterization of Geo-Net-PT. Tables adapted from [6].**(a)** Statistics of the Administrative Domain

Feature Type	#Features	
	absolute	relative / %
Postal Code	187,014	48.44
Street Segments	146,422	37.93
Settlement	44,386	11.50
Civil Parishes	4,260	0.93
Zone	3,594	0.08
Municipality	308	0.01
NUT	40	0.01
Districts	18	<0.01
Province	11	<0.01
Island	11	<0.01
Region	2	<0.01
Country	1	<0.01
Total	386,067	100.00

(b) Statistics of the Physical Domain

Feature Type	#Features	
	absolute	relative / %
Stream	2,421	42.65
Beach	588	9.83
Museum	507	8.93
Archaeological Site	414	7.29
Hotel	381	6.71
Natural Region	304	5.36
Castle	256	4.51
Spring	220	3.88
Historic Hamlet	217	3.82
Reservoir	90	1.59
Touristic Resource	84	1.48
Other	224	3.95
Total	5,676	100.00

(c) Relationships in Geo-Net-PT

Domain	<i>part-of</i>	<i>adjacent-to</i>
Administrative	386,431	33,051
Physical	389	2,404
Inter-Domain	2,752	0

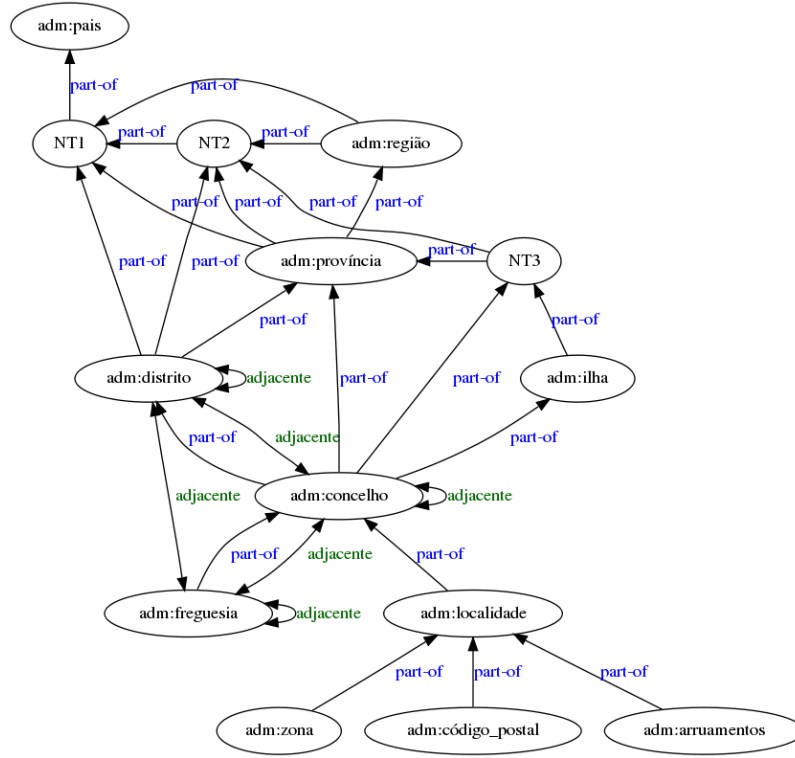


Figure 1: Graph of the possible relationships in the administrative domain of Geo-Net-PT.

Table 2: Feature types and their frequencies as of August 2010.

Feature Type		#Features	
		absolute	relative / %
Official	Town	21,976	93.59
	Postal Code	507	2.16
	County (or Municipality)	308	1.31
	Island	27	0.11
	State (or District)	20	0.09
	Country	1	<0.00
Informal	Point of Interest	494	2.10
	Suburb	95	0.40
	Airport	35	0.15
	Land Feature	10	0.04
	Colloquial	6	0.03
	Time Zone	3	0.01
Total		23,482	100.00

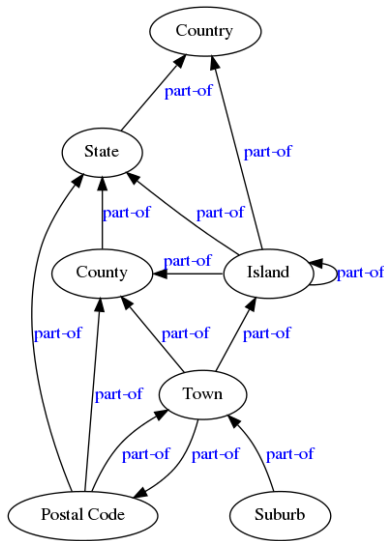


Figure 2: Graph of the possible relationships in GeoPlanetTM.

3 Methodology

YGP is mainly dedicated to administrative regions, containing both official and informal feature types. GNP contains administrative features as well as physical regions, but only official features are included in the ontology. As such, the alignment is restricted to official administrative features. In other words, we matched the GNP features with the YGP features whose type is listed in Table 3. Ideally, our alignment should make strong use of geospatial coordinates and names, but since only features of type district (**DST**), municipality (**CON**) and civil parish (**FRG**) have latitude and longitude coordinates, other information must be used to match features between the two ontologies. Therefore, GNP features are matched with YGP according to name, geospatial coordinates of the ancestors and type. The type concordance assures that the district named *Lisboa* does not get matched to the location named “*Lisboa*” in “*Viana do Castelo*”.

We proceed in two stages, each one responsible for finding different pairs of matching features. First we attempt to match postal codes and then other location types.

3.1 Stage 1: Postal Codes

The postal code system of the Portuguese territory has changed in 1998. Before that, postal codes were 4-digit codes that identified a major postal area. Nowadays, the system appends to the same 4 digits a hyphen and 3 digits that further specify postal routes.

GNP provides the full 7-digit Portuguese postal codes (e.g., “1100-254”), but YGP only has the first 4 digits. As such, we identify a match between `gnp:“1100-254”` and `ygp:“1100.”` This is not a match between corresponding entities, but expresses a hierarchical relation between the terms and, as it will

be seen later (Section “Concordance RDF File”), the resulting alignment takes into consideration the difference between postal codes and other feature types.

3.2 Stage 2: Other location types

On the second stage, we match the locations that are not postal codes. The workflow, which is detailed in the next sub-sections, is represented in Figure 3. It has six modules:

- **name similarity** gathers all (GNP, YGP) feature pairs having similar names;
- **type concordance** filters those pairs so that only the features of similar type pass to the next module;
- **distance** also filters the pairs so that only matched feature pairs which are physically close are retained;
- **manual pairs** adds manually matched pairs that other modules were unable to match automatically;
- **purge** removes pairs containing repeated features;
- **semantic propagation** determines whether the features in one of the discarded pairs in the purge step are semantically equivalent.

The remainder of this section explains how these modules work.

3.2.1 Name similarity

This module runs through every possible (GNP, YGP) feature pair, keeping only those whose names are similar. Because geographic names may not be identical among the ontologies, due to misspellings, we have implemented Algorithm 1 to detect similarity between the names. In this algorithm, we use the functions `name_to_words`, `levenshtein` and `misspelling`. The first converts the name to ASCII characters, replaces punctuation by spaces, spells numbers out and splits the entire string into space-separated tokens; `levenshtein` calculates the

Table 3: Type concordance between Geo-Net-PT and GeoPlanet™.

GNP type		YGP types
3-letter code	Name in English	
PAI	Country	Country
DST	District	State
ILH	Island	Island
CON	Municipality	County
CDP	Postal Code	Postal Code
FRG	Civil Parish	Town & Suburb
LOC	Settlement	Town & Suburb
ZON	Zone	Town & Suburb

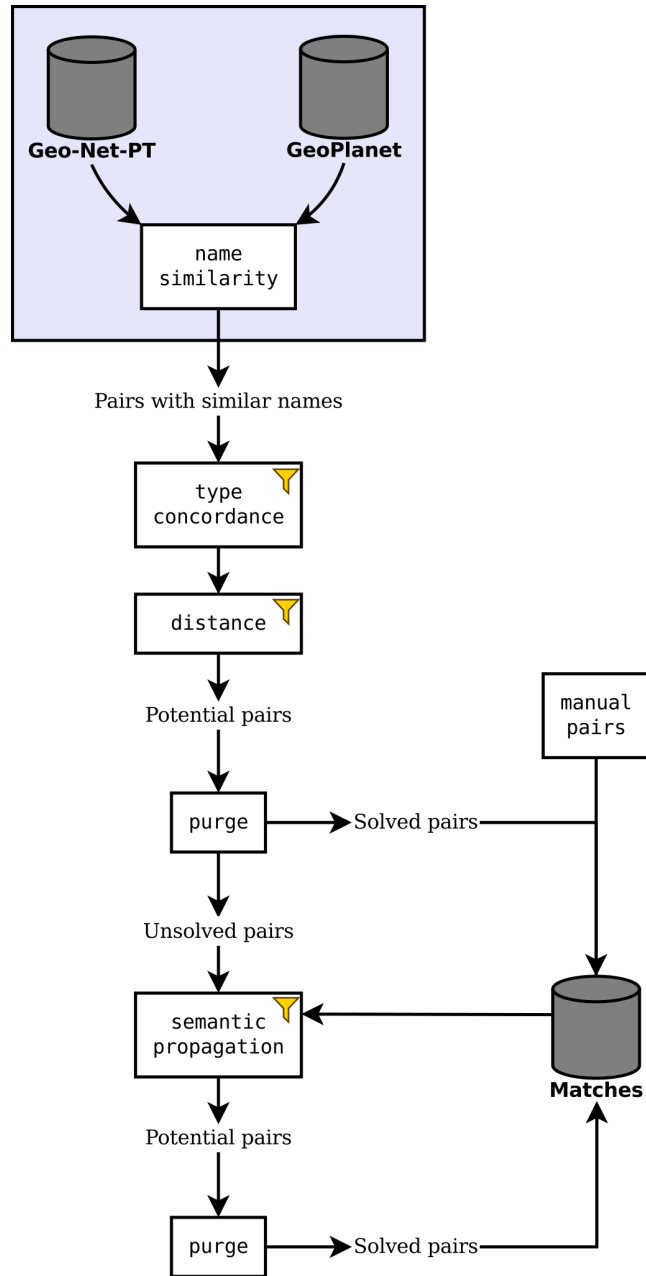


Figure 3: The workflow followed to create the alignment between Geo-Net-PT and GeoPlanet™.

Algorithm 1 `names_are_similar(gnp_name, ygp_name)`

```
seq1 ← name_to_words(gnp_name)
seq2 ← name_to_words(ygp_name)
diff_output ← compare seq1 to seq2 with diff
for all operation, batch1, batch2 ∈ diff_output do
  if operation = “delete” then
    if any word in batch1 has length > 1 then
      return False
    // Thus, “bondiosa a nova” and “bondiosa nova” are similar.

  else if operation = “insert” then
    if any word in batch2 has length > 1 then
      return False
    // Thus, “bondiosa nova” and “bondiosa a nova” are similar.

  else if operation = “replace” then
    if length(batch1) ≠ length(batch2) then
      return False
    for all (word1, word2) ∈ (batch1, batch2) do
      if number of features in GNP named gnp_name = 1 then
        if levenshtein(word1, word2) > 1 then
          return False
        else if not misspelling(word1, word2) then
          return False

  // At this point, all test succeeded for this diff operation.
  // Go on to the next operation.

return True
```

“edit distance” between two strings [4]; `misspelling` takes two words and returns True if the difference between the words is a single misspelling and False otherwise. By misspelling, we mean one of the following substitutions: “s” by “z”, “e” by “i”, “o” by “u” or the inverse. Since YGP names are given in the local language, in this case Portuguese, and GNP is multilingual and includes names in Portuguese, we do not expect any translation issues. However, the pair (“lisboa”, “lisbon”) is considered a special kind of misspelling, since “lisbon” is not localized in GeoPlanet™.

We also use the result of the `diff` algorithm, which determines which words are *deleted* from the first sequence, *inserted* in the second sequence, *replaced* from the first to the second sequence or *unchanged* between the two sequences.

Table 4 presents some names that this algorithm considers similar, along with some observations.

3.2.2 Type concordance

Several geographical features in Portugal have the same name. As an example, consider “Lagoa,” which is the name of 148 features in GNP. Of these, 42 have a type that can be aligned with YGP, namely 2 **CON**, 5 **FRG**, 34 **LOC** and

Table 4: Examples of (GNP, YGP) name pairs matched with Algorithm 1

GNP name	YGP name	Observations
Vale do André	Vale Andreu	“do” is ignored and since there is only one place with the name “Vale do André” in GNP (gnp:138940), “André” and “Andreu” are compared with a Lavenshtein algorithm.
Vale d’El Rei	Vale de El-Rei	Punctuation is replaced with spaces; “d” and “de” are ignored.
Bernalfor	Bernalflor	There is exactly one GNP place named “Bernalfor”, thus these words are compared with a Lavenshtein algorithm.
Carviçais	Carviçaes	The difference is an “i” that becomes an “e”.
Bondiosa a Nova	Bondiosa Nova	The deleted word, “a”, has length 1.
7 Casas	Sete Casas	Numbers are converted to their name.

1 **ZON**. In YGP, there are 24 features with the same name, 2 **Counties** and 22 **Towns**. This creates $42 \times 24 = 1008$ pairs. According to Table 3, only a subset of these are considered: the 2 **CON** are paired with the 2 **Counties** and the 5 **FRG**, 34 **LOC** and 1 **ZON** are paired with the 22 **Towns**, for a total of 884 possible pairs.

The first five rows in Table 3 are obvious: for instance, districts pair up with districts (which, in YGP, are denominated States). The last six rows, however, need further explanation.

YGP does not have the concept of civil parish, represented in GNP by the type **FRG**. Therefore, we can find GNP features of type **FRG** scattered among **Towns** and **Suburbs** in YGP. Likewise, GNP does not have a clear concept of town, represented in YGP by the type **Town**. Therefore, we can find YGP features of type **Town** scattered among features with type **LOC** or **ZON**. Since **Suburbs** are *part-of* **Towns**, they must also be matched to these GNP types. Some examples can be found in Table 5.

Table 5: Examples of matches between Geo-Net-PT features of type **FRG**, **LOC** and **ZON**. For each of the examples, the number of features in both ontologies with that name is unique, which means that the features could not have been matched at all if the corresponding type concordance did not exist.

GNP type	YGP type	Feature names
FRG	Town	Macinheta da Seixa
FRG	Suburb	Aldoar
LOC	Town	Fartaria
LOC	Suburb	Miraflores
ZON	Town	Casal da Misarela
ZON	Suburb	Olivais Norte

3.2.3 Distance

The next step is to filter the pairs by distance: it is sensible to expect each of the 2 **CON** features in GNP to be matched with one of the 2 **Counties**, but, as we have seen, each **CON** feature is paired with both **Counties**, for a total of 4 pairs. To disambiguate, we use the physical distance between the features of each pair and discard those which are too distant. In Figure 4, we show the distances of the 4 pairs. Since the features in the pairs highlighted with the bold lines are physically near and have compatible types, the two pairs are kept; the other two pairs, represented with the dotted lines, have distances higher than 1500 km, and are discarded. Hence, the distance module acts as a filter that only keeps the pairs for which the calculated distance is small enough.

There are two issues in this method. First, not all GNP features have geospatial coordinates, and second, we need to define which distances are small enough to keep a given pair.

In GNP, features of type **DST**, **CON** and **LOC** have geospatial coordinates, and **FRG**, **LOC** and **ZON** features are *part-of* exactly one **CON**. Hence, for these types, we use the coordinates of the **CON** to which they belong in order to calculate the distance between the features of the pair. To decide if a pair should be kept, this module uses the radius of features, which it approximates to the radius of the flat circle with an area equal to the area of the feature: $r \simeq \sqrt{A/\pi}$. The mean radius of a type is calculated as the average of the radius of all features of that type. Only GNP has information about the area of features, and only for features of type **DST**, **CON** and **FRG**. The types **LOC** and **ZON** are assumed to have the same mean radius as **FRG**.

Let T be the type of the GNP feature of a pair. Then, the pair is kept if:

1. the GNP place has coordinates, and the distance is not higher than the mean radius of type T ; or
2. the GNP place does not have coordinates, and the distance between the parent of the GNP place and the YGP place is not higher than the radius of the parent plus the mean radius of T .

Figure 5 illustrates the reasons behind these rules. In the first case, depicted in Figure 5a, we use the mean radius of T as the threshold to accommodate for possible inconsistencies resulting from the different sources of the information.

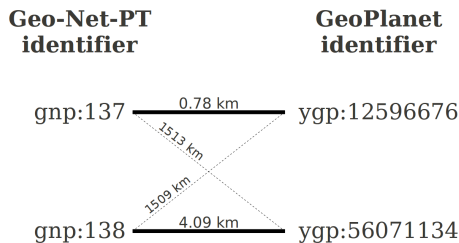


Figure 4: The distance between the Geo-Net-PT and GeoPlanet™ features named “Lagoa”. Types are **CON** and **County** respectively. The bold lines show the pairs that are kept.

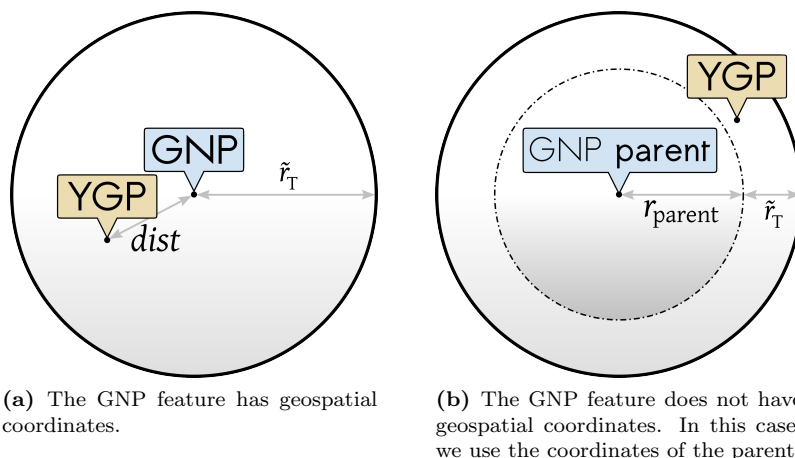


Figure 5: The distance between a GNP feature and a YGP feature must be less than or equal to a threshold that is defined according to the type of the GNP feature. (a) When the GNP feature has coordinates, the threshold is the mean radius of the GNP feature type, \tilde{r}_T ; (b) When the coordinates of the GNP feature are not known, the threshold is the radius of the parent of the GNP feature, r_{parent} , plus the mean radius of the GNP feature, \tilde{r}_T . In the two represented cases, the distance between the features in the pairs is smaller than the threshold.

In Figure 5b, the GNP feature does not have geospatial coordinates, which means that there is no way to calculate the distance between the features in the pair. In this case, we use the ancestor of the GNP feature of type **CON** to calculate a distance, which means that the threshold must be higher than simply the mean radius of the GNP feature. In the ideal case, the GNP and YGP features would have the exact same coordinates and features of the type **CON** would be perfect flat circles; then, we could simply test the distance against the radius of the GNP feature’s ancestor. However, since there are inconsistencies between the two ontologies and feature shapes are not circles, we increase the threshold by the mean radius of T to compensate for these effects. Another reason to use such a threshold is that geospatial coordinates are expressed in different geospatial data (ETRS89 datum for GNP and WGS84 for YGP).

Continuing with the example of the previous section, **gnp:398552** (“Lagoa”, **LOC**) does not have geospatial coordinates, and as such we need to use the coordinates of its ancestor of type **CON**, **gnp:334** (“Vila Pouca de Aguiar”). According to the discussion above, the threshold is $r_{\text{gnp:334}} + \tilde{r}_{\text{LOC}} = 11.80 + 2.22 = 14.02$ km. The matching YGP feature must be one of the 22 **Towns** found in the previous section; the only one of these whose distance to **gnp:334** is smaller than 14.02 km is **ygp:742383**, making the two a matching pair.

On the other hand, when the same analysis is applied to **gnp:48092**, another one of the features of type **LOC** that are named “Lagoa,” we find two YGP features: **ygp:29386127** and **ygp:55867935**. This means that this module is unable to find a unique match in YGP for that GNP term.

3.2.4 Manual pairs

Some pairs have to be manually added to the alignment, including the top-level pair (`gnp:“Portugal”`, `ygp:“Portugal”`). `gnp:“Portugal”` does not have coordinates or parents, which means that it is filtered out by the `distance` module. Other manually added pairs include pairs of the type **ILH** (**I**sland in YGP), which were discarded by the `name similarity` module since the names differ significantly, as it happens, for instance, between “Ilha do Faial” and “Fayal Island”, the names for an Island in the Azores in GNP and YGP, respectively.

Finally, two pairs that break the type concordance were also manually added. The Portuguese archipelagos were once considered districts of the country, and, as such, YGP has the terms `ygp:“Azores”` (**State**) and `ygp:“Madeira”` (**State**). These two regions are no longer categorized as Districts, but as Autonomous Regions, and the YGP features were matched to `gnp:“Região Autónoma dos Açores”` (**NT3**) and `gnp:“Região Autónoma da Madeira”` (**NT3**), respectively.

3.2.5 Purge

After running the previous modules, we get a series of (GNP, YGP) feature pairs, representing possible matches in the alignment. There is, however, no guarantee that all GNP features and all YGP features are unique: consider the previous example, where `gnp:48092` (“Lagoa”, **LOC**) is paired with two distinct YGP features. Other features with the same or a similar name may be close to one another, resulting in several pairs sharing the same GNP or YGP feature. These potential pairs should not figure in the final alignment, because there is no way to disambiguate them.

To enhance the flexibility of this alignment, we have partitioned all matches such that in each partition all pairs have the same GNP type and the same YGP type. We ran the `purge` method to each partition individually. Thus, if there are two YGP **Towns** matched to a single GNP feature of type **LOC**, both matches must be discarded, but if the repeated GNP feature is matched to a single **Town** and a single **Suburb**, we keep both matches.

3.2.6 Semantic propagation

At this stage in the workflow, there are already many pairs in the alignment. However, some features remain to be matched. To improve the recall of the alignment, we used semantic information about each feature. For every pair, we extracted the ancestry of the GNP and YGP features, keeping only those ancestors with a type present in Table 3. By querying the pairs already selected for the alignment, we determined whether the two ancestries were the same.

Reusing the previous example, `gnp:48092` was paired with two YGP features, which resulted in its exclusion from the alignment. However, when this module runs over the GNP feature, it finds its ancestry to be the same as the ancestry of `ygp:55867935`, as shown in Table 6. The pairs on that table (except the first row) were found in the previous stages of the matching algorithm. This means that the new pair is redrawn from the discarded pairs and is kept as another match.

Since this step can also lead to features appearing in more than one of the selected pairs (which would happen here if the second YGP feature was also mapped to the Braga Municipality), we run the `purging` method again.

Table 6: The ancestry of the settlement “Lagoa” in the Braga municipality. The ancestors in both ontologies are retrieved. Each row represents a match already determined in previous steps, except for the first one, which is the pair being aligned.

Geo-Net-PT				Yahoo! GeoPlanet™		
ID	Name	Type		WOEID	Name	Type
48092	Lagoa	LOC	—	55867935	Lagoa	Town
64	Braga	CON	—	12596498	Braga	County
3946	Braga	DST	—	2346564	Braga	State
418745	Portugal	PAI	—	23424925	Portugal	Country

3.3 Validation

The output of the methodology herein described is a set of (GNP, YGP) feature pairs. We validate this set of pairs based on the relationships they have on both ontologies.

For each pair, we seek to determine whether a *significant* ancestor of the GNP feature is part of the ancestry of the YGP feature. As such, the validation process must start in a manually validated root. The pairs for the districts (found automatically) and for the islands (assigned manually) act as roots.

For pairs where the GNP feature has type **CON**, we retrieve the ancestor of type **DST** or, in case one does not exist, the ancestor of type **NT3** (see Section 3.2.4, “Manual pairs”). By querying the already validated pairs, we convert this ancestor to a YGP feature and, if this is present in the ancestry of the YGP feature of the pair, the pair is validated. All features of type **CON** were validated. For features of type **FRG**, **LOC** and **ZON**, we retrieve the ancestor of type **CON** or **ILH** and run the same process.

For example, the term `gnp:91` (“Coimbra”, **CON**) matches `ygp:12596482` (“Coimbra”, **County**). Since they are both part of the “Coimbra district” (terms `gnp:3949` and `ygp:2346567`), the match is valid. On the other hand, term `gnp:24764` (“Goujeva”, **LOC**) is matched to `ygp:56041720` (“Gougeva”, **Town**), but the GNP feature is located in the “Santa Maria da Feira” municipality and the YGP feature in the “Vila Nova de Gaia” municipality, which are not paired. This invalidates the pair.

4 Results

Table 7 shows the number of matches obtained after the stage 1 of the matching process. We have aligned almost all postal codes from GNP (98.5%) to a postal code in YGP. The number of unique YGP postal codes used for these matches is 507, which represents 100% of the postal codes in this ontology.

The results of stage 2 are shown in Table 8. It is visible that YGP has 2 more districts than GNP. The extra two are “Azores” and “Madeira,” which are no longer considered districts (as discussed in Section 3.2.4, “Manual pairs”). Features of type **CON** are all correctly matched between the two ontologies. There are more islands in YGP than in GNP, because YGP considers small uninhabited islands and GNP only considers the 11 main islands in the two

Table 7: Summary of alignment of postal codes (stage 1 of the matching process).

Types		Totals		Matches
GNP	YGP	GNP	YGP	
CDP	Postal Code	187,014	507	184,240

Table 8: Summary of the alignment of other types (stage 2 of the matching process).

Types		Matches	Validated
GNP	YGP		
DST	State	18	18
CON	County	308	308
ILH	Island	11	11
FRG	Town	3,503	3,495
FRG	Suburb	26	26
LOC	Town	14,236	14,027
LOC	Suburb	4	4
ZON	Town	711	668
ZON	Suburb	31	31
Total		18,848	18,588

archipelagos. These 11 islands were manually paired and thus automatically considered valid. Each GNP feature of type **FRG**, **LOC** and **ZON** matches at most one YGP feature. This means that no GNP feature is repeated in the alignment. On the other hand, some **Towns** and **Suburbs** from YGP have more than one associated GNP feature.

The statistics in Tables 9 to 11, particularly those for the types **FRG**, **LOC**, **ZON**, **Town** and **Suburb**, represent the core of our work. We can see that, overall, in GNP, only 34.9% of these features were matched (recall). The recall relative to YGP is much higher, 71.0%. Thus, the smaller recall in GNP is due GNP being more comprehensive than YGP. Other reasons for this difference include the errors associated with the geospatial coordinates of each ontology and the errors that can arise in the spelling of names. Nevertheless, our method was able to match 82.7% of all **FRG**, which, together with the features of type **CON** and **DST** contribute to a wide coverage of the administrative division of Portugal.

Of all the matches, 260 (~1.3%) did not pass the validation process. Many of these are due to the fact that features may be incorrectly annotated in each ontology. For instance, `gnp:53318` (“Serafão”, **LOC**) is *part-of* `gnp:“Fafe”` (**CON**) but it is matched to `ygp:748678` (“Serafão”, **Town**) which is *part-of* `ygp:“Guimarães”` (**County**). GNP also has the feature `gnp:1342` (“Serafão”, **FRG**), which is located in `gnp:“Fafe”`. This seems to suggest that the YGP relation is incorrect.

The number of matches achieved after each step of the workflow are given

Table 9: Summary results for the Geo-Net-PT features.

GNP Type	Totals	Matches	Validated
DST	18	18	18
CON	308	308	308
ILH	11	11	11
FRG	4,260	3,529	3,521
LOC	44,386	14,240	14,031
ZON	3,594	742	699
Total	52,577	18,848	18,588

Table 10: Summary results for the GeoPlanetTM features. Repeated features (those that are matched to more than one Geo-Net-PT feature) are not considered.

GNP Type	Totals	Matches	Validated
State	20	18	18
County	308	308	308
Island	27	11	11
Town	21,976	15,826	15,622
Suburb	95	45	45
Total	22,426	16,208	16,004

Table 11: Precision and recall for the alignment. Precision measures the validity of each match, recall measures the coverage relatively to the size of each ontology.

Ontology	Type	Precision	Recall	F-measure
GNP	DST	1.000	1.000	1.000
	CON	1.000	1.000	1.000
	ILH	1.000	1.000	1.000
	FRG	0.998	0.827	0.904
	LOC	0.985	0.316	0.478
	ZON	0.942	0.194	0.322
	FRG/LOC/ZON	0.986	0.349	0.516
YGP	Total	0.986	0.354	0.521
	State	1.000	0.900	0.947
	County	1.000	1.000	1.000
	Island	1.000	0.417	0.589
	Town	0.987	0.711	0.827
	Suburb	1.000	0.474	0.643
	Town/Suburb	0.987	0.710	0.826
Total	0.987	0.714	0.829	

Table 12: Partial results after each of the steps in the workflow. The number of pairs that figure in the final alignment are written in *italic*.

Workflow Module Name	#pairs
name similarity	218,686
type concordance & distance	21,127
purge	<i>17,702</i>
manual pairs	<i>15</i>
semantic propagation	2,036
purge	<i>1,131</i>
Total	<i>18,848</i>

in Table 12, which represents partial results for stage 2. The number of pairs decreases at each step, reflecting that the methodology consists in applying a series of filters that discard wrong or unsolved pairs.

4.1 Concordance RDF File

The most obvious choice for aligning the features in each ontology would be *owl:sameAs*. However, the semantics of the associated relation is too strong, and we were unable to verify if every match represents two identical features in semantic terms. The difference in the types of both ontologies, particularly the fact that YGP does not make a distinction between settlements and

```

<!ENTITY gnpt02 "http://xldb.di.fc.ul.pt/xldb/publications/2009/10/geo-net-pt-02#">
<!ENTITY gplnt "http://where.yahooapis.com/v1/place/">

<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:skos="http://www.w3.org/2004/02/skos/core#"
  xmlns:dc="http://purl.org/dc/terms/"
  xmlns:gplnt="&gplnt;"
  xmlns:gnpt02="&gnopt02;" >

  <rdf:Description rdf:about="&gnpt02;172208">
    <skos:broaderMatch rdf:resource="&gplnt;24553912"/>
  </rdf:Description>
  [...]

  <rdf:Description rdf:about="&gnpt02;moita-AF401996">
    <skos:closeMatch rdf:resource="&gplnt;743651"/>
  </rdf:Description>
  [...]

</rdf:RDF>

```

Figure 6: A snippet of the alignment, in RDF format.

their respective civil parishes, also invalidates the use of that relation. Finally, aligning ontologies with the *owl:sameAs* relation can result in unwanted results, as shown in the SKOS documentation (see <http://www.w3.org/TR/2009/REC-skos-reference-20090818/#L4858> [10]). For these reasons, we choose to use *skos:closeMatch* instead, which reflects more accurately that the two concepts “can be used interchangeably in some information retrieval applications” [10]. For postal codes, however, the aligned entities are not identical, but are instead hierarchically related. Thus, we use *skos:broadMatch*.

The alignment is published online in RDF/XML format as a supplemental information to this document at hdl.handle.net/10455/6677. Figure 6 contains a snippet of the file, showing the XML header and two examples of matches, one for each stage of the workflow:

4.2 SPARQL queries

Geo-Net-PT has a SPARQL [9] endpoint that can be used to query the ontology, at http://xldb.fc.ul.pt/wiki/Geo-Net-PT_02_SPARQL_endpoint [2]. We added the alignment to the endpoint, allowing the inclusion of GeoPlanet™ features in queries:

```
SELECT ?parishlabel, ?beachlabel WHERE {
  ?geonet skos:closeMatch <http://where.yahooapis.com/v1/place/12596520> .
  ?geonet gnpt:hasPart ?parish .
  ?parish gn:type gnpt02:freguesia-ATFRG .
  ?beach gnpt:isLocatedOn ?parish .
  ?beach gn:type gnpt02:praia-PTPRA .
  ?parish rdfs:label ?parishlabel .
  ?beach rdfs:label ?beachlabel .
}
```

The query above searches for GNP features of type **Praia** (beach) that belong to some civil parish in the Municipality of Faro (*ygp:12596520*). It retrieves the name of the civil parish and the name of the beach using the GNP relation *gnpt:isLocatedOn*, which connects the physical and the administrative domains of the ontology [6]. Labels contain the type of the feature are between parenthesis. The results are shown in Table 13.

5 Discussion and Conclusions

The work described here has culminated in the production of an alignment that contains a concordance between Geo-Net-PT and Yahoo! GeoPlanet™.

Table 13: The results of the SPARQL query

?parishlabel	?beachlabel
Sé (Freguesia)	Ilha de Faro (Praia)
Sé (Freguesia)	Ilha da Barreta (Praia)
Sé (Freguesia)	Ilha da Culatra (Praia)
Sé (Freguesia)	Ilha do Farol (Praia)

If used by Yahoo on its geographical ontology, it will enhance the visibility of Geo-Net-PT, which in the long term would mean a better feedback from users, coupled with the improvement in quality that this feedback would entail. Another possible improvement would be the validation of the geospatial coordinates in Geo-Net-PT, since they have origin in different geospatial coordinate systems and projections.

As mentioned in the introduction, this alignment contributes to a better annotation of information resources, since it partially removes the ambiguity of annotating resources with one of these ontologies. When news, models or other computational object are annotated with a Yahoo! GeoPlanet™ term matched to some Geo-Net-PT feature, the annotation becomes richer, since this ontology provides more detail than the Portuguese territory described under GeoPlanet™.

The alignment produced is relatively complete, containing almost all the administrative divisions of Portugal (districts, municipalities and civil parishes) as well as a high number of other towns.

In conclusion, we have produced an alignment between a broad geospatial ontology, Yahoo! GeoPlanet™, and a narrow geospatial ontology, Geo-Net-PT, using information about the names and geospatial coordinates of the features in the matching process. The results were stored in an RDF file that can be queried through a SPARQL endpoint or downloaded from <http://hdl.handle.net/10455/6677>. The whole Geo-Net-PT ontology can also be downloaded from the Linguatca website as a PostgreSQL database dump file.

6 Acknowledges

This work was supported in part by FCT (Portuguese research funding agency) for its LASIGE Multi-annual support, GREASE-II project (grant PTDC/EIA/73614/2006). We would like to thank Eddie Babcock from Yahoo! Geo Technologies for the valuable advices and comments on the alignment produced and Francisco J. Lopez-Pellicer and Nuno Cardoso for their important support on Geo-Net-PT and RDF Specifications.

References

- [1] T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. *Scientific American*, 284(5):28–37, 2001.
- [2] Nuno Cardoso and Mário J. Silva. A GIR Architecture with Semantic-flavored Query Reformulation. In *6th Workshop on Geographic Information Retrieval*, Zurich, Switzerland, 18-19 February 2010.
- [3] J. Euzenat and P. Shvaiko. *Ontology Matching*. Springer-Verlag New York Inc, 2007. ISBN 978-3-540-49611-3.
- [4] V. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710, 1966.
- [5] L. Lopes, F. Silva, F. Couto, J. Zamite, H. Ferreira, C. Sousa, and M. Silva. Epidemic Marketplace: An Information Management System for Epidemi-

ological Data. *Information Technology in Bio-and Medical Informatics, ITBAM 2010*, pages 31–44, 2010.

- [6] Francisco J. Lopez-Pellicer, Marcirio Chaves, Catarina Rodrigues, and Mário J. Silva. Geographic Ontologies Production in Grease-II. Technical Report TR 09-18, University of Lisbon Faculty of Sciences LASIGE, November 2009.
- [7] M. Luczak-Rösch and R. Heese. Linked Data Authoring for Non-Experts. In *Linked Data on the Web Workshop*, 2009.
- [8] W3C. Resource Description Framework (RDF) Model and Syntax, 1999. <http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/>.
- [9] W3C. SPARQL Query language, 2008. <http://www.w3.org/TR/rdf-sparql-query/>.
- [10] W3C. SKOS Simple Knowledge Organization System Reference, 2009. <http://www.w3.org/TR/2009/REC-skos-reference-20090818/>.
- [11] Yahoo. Yahoo! GeoPlanet™, Accessed July 2010. <http://developer.yahoo.com/geo/geoplanet>.