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

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

DI–FCUL–TR–2010–5

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! GeoPlanet™ 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 7, 2010

**Abstract**

**Background:** Geo-Net-PT is a geospatial ontology representing the Portuguese territory and the relations between the several locations within in. Yahoo! GeoPlanet™ is a geospatial ontology that covers the whole world. To diminish the effects of repeated information, we propose an alignment between the administrative part of these two ontologies based on name similarity and physical closeness.

**Results:** After running the matching process, 16,814 matches were found, corresponding to 33% of the considered features in Geo-Net-PT and to 75% in Yahoo! GeoPlanet™. 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 enables the mapping of annotation in documents and other resources from GeoPlanet™, a broad ontology, to Geo-Net-PT, a specific ontology for Portuguese geography.

## 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 [6], much in the style advocated by the Semantic Web [1], but they also allow the use of automatic reasoning over the geographical locations.

The geospatial information on Earth is so wide that it is beneficial to have ontologies spanning ranges 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! GeoPlanet$^{TM}$ [12] (YGP), an ontology that covers places from all around the world; Geo-Net-PT [7, 9] (GNP) is an example of the second case, since it is a geospatial ontology covering only the Portuguese territory [7].

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! GeoPlanet$^{TM}$ refer to the same entity, removes the ambiguity and a search engine could actually retrieve that same document even if the query term was the term from Yahoo! GeoPlanet$^{TM}$.

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), 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, which comes particularly useful in a Semantic Web [1] world. LinkedData [8] 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 in triples in an RDF file. In the same spirit, we created an RDF [4] file containing all the pairs found for the matching of the two ontologies.

These triplets link features in GNP to features YGP through the use of SKOS (Simple Knowledge Organization System) properties. SKOS is a system that develops a standard way to represent knowledge using the Resource Description Framework (RDF) [11]. 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 statistical results 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 [7, 12]. 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 at Linguateca in `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 in `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 [7]. 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 GeoPlanet™

Yahoo! GeoPlanet™ [12] is a world coverage geographic ontology. Each place in this ontology is identified by a unique identifier dubbed *Where On Earth Identifier* (WOEID). In the Portuguese part of this ontology, each place can have one of the types in Table 2.

In that table, there is 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 used the parent-child relationships, defined as "the direct inferiors to a given place." This is equivalent to a *has-part* relation, the inverse of *part-of*. In this version of GeoPlanet™, 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 related to one another through *part-of* relationships in a hierarchical, as shown in Figure 2.

In the beginning of the work, we retrieved all information from YGP concerning the Portuguese territory. To do this, we queried Yahoo!'s GeoPlanet™ web service in order to retrieve ygp:23424925 "Portugal", its children and the children of all its descendants, recursively, until no more features were found.

**Table 1:** Characterization of Geo-Net-PT. Tables adapted from [7].

**(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** | ***part-of*** | ***adjacent-to*** |
|---|---|---|
| Administrative | 386,431 | 33,051 |
| Physical | 389 | 2,404 |
| Inter-Domain | 2,752 | 0 |

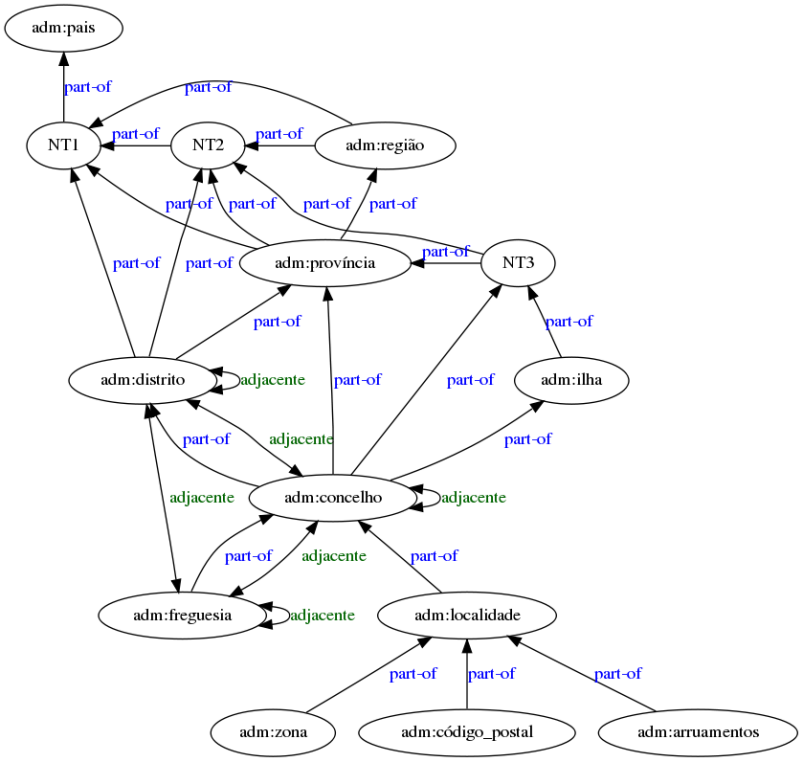



**Figure 1:** Graph of the possible relationships in the administrative domain of Geo-Net-PT.

**Table 2:** Feature types in GeoPlanet$^{TM}$. These values were retrieved from the local copy of GeoPlanet$^{TM}$.

| **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 |

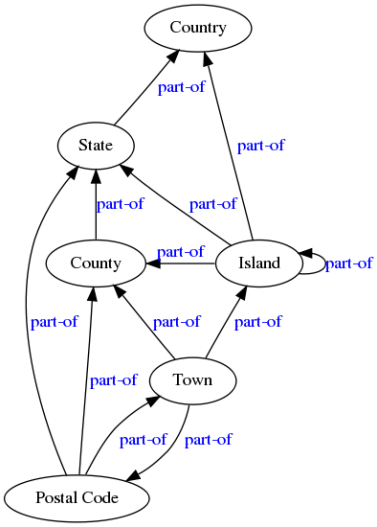



**Figure 2:** Graph of the possible relationships in GeoPlanet$^{TM}$.

# 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. This fact led us to restrict the alignment only to official administrative features. As such, we matched the GNP features having a type present in Table 4. 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 were 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".

Our work was divided in two different stages, each one responsible for finding different pairs of matching features.

## 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 uses the same 4 digits accompanied by a hyphen and 3 digits that further specify postal routes.

All the postal codes have been matched as follows. GNP provides the full 7-digit Portuguese postal codes (e.g., "1100-254"), but YGP has only the first 4 digits. As such, we created 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 will be seen later (section "Concordance RDF File"), the resulting alignment makes takes into consideration the difference

between postal codes and other feature types.

## 3.2 Stage 2: Other locations

On the second stage, we matched 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**: this module gathers all (GNP, YGP) feature pairs where the names of the features are similar;
- **type concordance**: this module filters those pairs so that only features with similar type pass to the next module;
- **distance**: this module also filters the pairs so that only features which are close remain;
- **manual pairs**: a selected number of pairs had to be matched because the other modules were unable to do it automatically;
- **purge**: this module removes pairs that contain repeated features;
- **semantic propagation**: this module determines whether the features in one of the pair discarded in the purge step are semantically equivalent.

Each of the next subsections explains how these modules work.

### 3.2.1 Name similarity

This module runs through every possible (GNP, YGP) feature pair, keeping only those whose name are similar. Because names may not (and often they are not) exact string matches from one ontology to the other, due to misspellings, we have implemented a specific algorithm to detect similarity between the names (Algorithm 1).

In this algorithm, we use the functions `name_to_words`, `levenshtein` and `misspelling`. The first reduces the name to ASCII characters, replaces punctuation by spaces and numbers by their portuguese name and splits the entire string into a sequence of space-separated tokens; `levenshtein` calculates one of the possible distances between two strings, usually called the "edit distance" [5]; `misspelling` takes two words and returns True if the difference between the words is a misspelling problem and False otherwise. By misspelling, we mean one of these six substitutions: "s" to "z", "e" to "i", "o" to "u" or the reverse. The pair ("lisboa", "lisbon") is considered a special kind of misspelling, since this is the only word that is not localized in GeoPlanet™.

We also used the result of the `diff` algorithm, which exposes batches of words from each sequence such that they are either *deleted* from the first sequence, *inserted* in the second sequence, *replaced* from one sequence to the other or *unchanged* between the two sequences.

Table 3 presents some names that this algorithm considers similar, along with some observations.

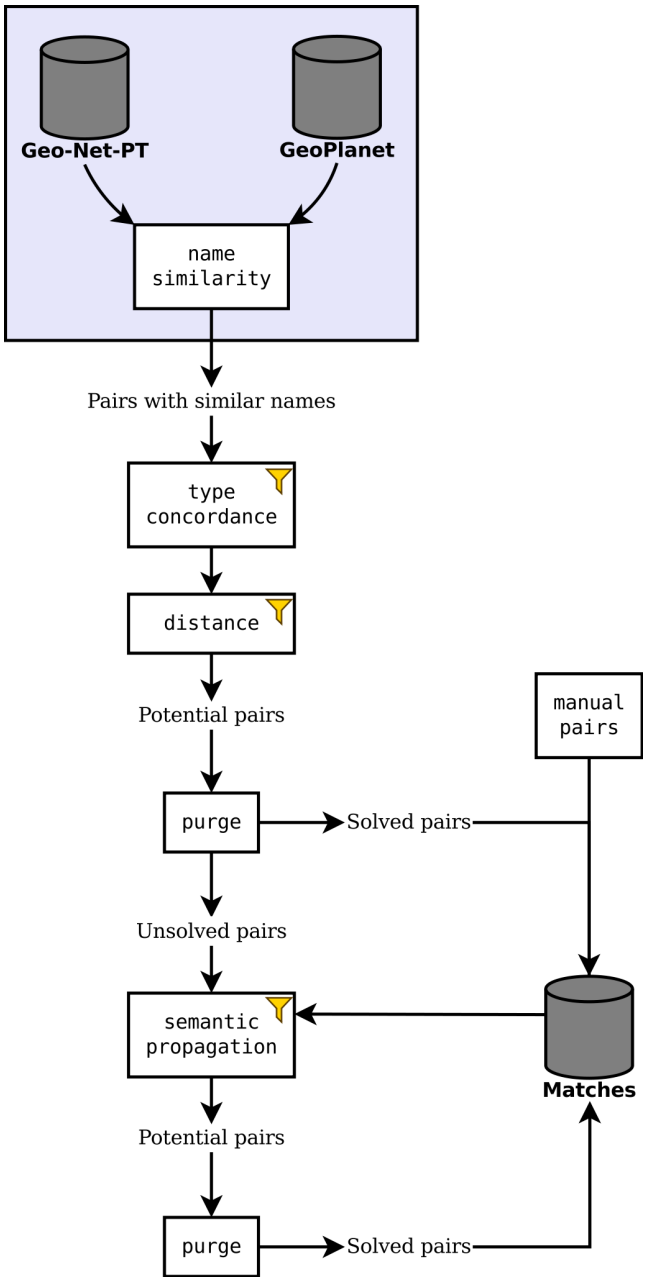



**Figure 3:** The workflow followed to create the alignment between Geo-Net-PT and GeoPlanet$^{TM}$.

**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
```

**Table 3:** Several similar (GNP, YGP) name pairs are present in this table, along with the result of the algorithm explained

| **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. |

**Table 4:** Type concordance between Geo-Net-PT and GeoPlanet$^{TM}$. Only pairs of features with matching types were considered to match the two ontologies.

| **GNP type** | | **YGP type** |
|---|---|---|
| 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 |
| FRG | Civil Parish | Suburb |
| LOC | Settlement | Town |
| LOC | Settlement | Suburb |
| ZON | Zone | Town |
| ZON | Zone | Suburb |

#### 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 1 **ZON**. In YGP, there are 24 features with the same name, 2 **Counties** and 22 **Towns**. This creates $42 \times 24 = 1008$ pairs. Based on type, we can reduce the number of potential pairs. According to Table 4, 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. It is worth noticing here that, because YGP only has 24 features with this name, and we are trying to match features in both ontologies in one-to-one links, the maximum number of pairs that can be formed is 24.

The first five rows in Table 4 were trivial to determine: 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 **LOC** and **ZON**. Since **Suburbs** are *part-of* **Towns**, they must also be matched to these GNP types. Examples can be found in Table 5.

#### 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, there are 4 pairs. To disambiguate, we use the physical distance between the two features of each pair and discard those pairs that are

**Table 5:** Examples of matches between Geo-Net-PT features of types **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 | Name of the features |
|---|---|---|
| FRG | Town | Macinheta da Seixa |
| FRG | Suburb | Aldoar |
| LOC | Town | Fartaria |
| LOC | Suburb | Miraflores |
| ZON | Town | Casal da Misarela |
| ZON | Suburb | Olivais Norte |

too distant. In Figure 4, we show the distance of these 4 pairs. The fact that the pairs highlighted with the bold lines are physically near plus the compatibility of the type are sufficient to keep these two pairs; the other two pairs, represented with the dotted lines, have distances higher than 1500 km, and are discard.

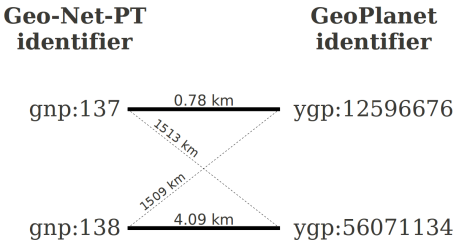



**Figure 4:** The distance between the Geo-Net-PT and GeoPlanet$^{TM}$ features of type CON and County respectively. The bold lines show the pairs that are kept.

This method has two problems. 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, only features of type **DST**, **CON** and **LOC** have geospatial coordinates, but the types **FRG**, **LOC** and **ZON** are all *part-of* exactly one **CON**. Thus, for these types, we used the coordinates of the **CON** to which they belong in order to calculate the distance between the features of the pair.

Thus, this module acts as a filter that keeps only pairs for which the distance calculated for the pair is small enough. To decide whether the pair should be kept, the filter uses the radius of features. The radius of a feature is approximated to the radius of the flat circle that has an area equal to the area of that 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** or **FRG**. The types **LOC** and **ZON** are assumed to have the same mean radius as **FRG**.

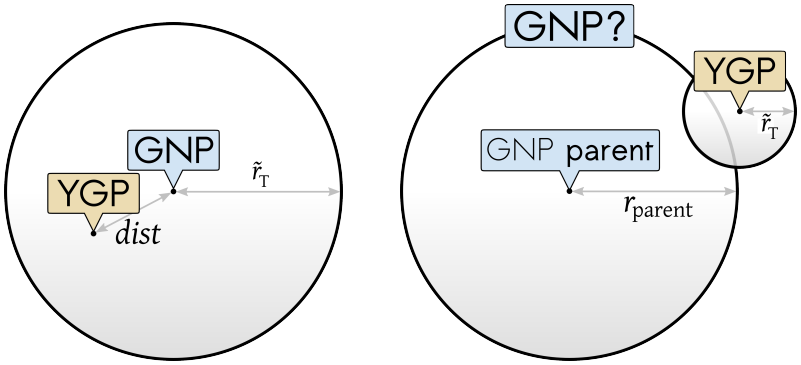
Let $T$ be the type of the GNP feature of a pair. Then, the pair is kept if:

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. In the second case, Figure 5b, the GNP feature does not have geospatial coordinates, which means that there is no way to calculate have the distance between the features in the pair. In these cases, we use an ancestor of type **CON** of the GNP feature 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, as above, inconsistencies between the two ontologies and the fact that features are not circles made us increase the threshold by the mean radius of $T$.




**(a)** The GNP feature has geospatial coordinates.

**(b)** The GNP feature does not have geospatial coordinates. In this case, we use the coordinates of the parent

**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 type of the GNP feature, $\tilde{r}_T$; (b) When the whereabouts of the GNP feature is unknown, the threshold is the radius of the parent of the GNP feature, $r_{\text{parent}}$, 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.

Consider the following example, falling under case 2 above. This example continues the one started in 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 **Tows** found in the previous section; the only one of these whose distance to gnp:334 is smaller than 14.02 km is ygp:742383, and this pair formed is kept.

On the other hand, when the same analysis is applied to gnp:48092, another one of the 34 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 single match for this GNP term.

### 3.2.4 Manual pairs

Some pairs had to be manually added to the alignment. The most important one was the pair (gnp:"Portugal", ygp:"Portugal"). This is because the GNP feature does not have coordinates and does not have parents, which means it is filtered out by the `distance` module. Other pairs manually added include trivial pairs of the type **ILH** (**Island** in YGP), which were discarded by the `name similarity` module since the names differ significantly, as happens, for instance, between "Ilha do Faial" and "Fayal Island", which are the names of an Island in 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 removed, 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 4. By querying the pairs already selected for the alignment, we determined whether the two ancestries were the same.

**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 we are trying to align.

| **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 |

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 possible match.

Since this step could 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 ran the purging method again.

## 3.3 Validation

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

For each pair, we sought to determine whether a *significant* ancestor of the GNP feature was part of the ancestry of the YGP feature. Districts were validated manually, since there are only 18 of those in Portugal.

For pairs where the GNP feature has type **CON**, we retrieved 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 converted 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 retrieved the ancestor of type **CON** or **ILH** and ran 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.

The pair (gnp:"Portugal", ygp:"Portugal") and the other manual pairs, were used as root of this validation process.

**Table 7:** Summary of the results achieved for the postal code alignment (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 results achieved for the other types alignment (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 |

# 4 Results

Table 7 shows the results 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 results of stage 2 are shown in Table 8. By analysing these results, we can see 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. The island type has more instances in YGP than GNP; this is because YGP considers small uninhabited islands and GNP only considers the 11 main islands in the two archipelagos. These 11 islands were manually paired and thus automatically validated. Each GNP feature of type **FRG**, **LOC** and **ZON** has been matched to 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 results displayed in Tables 9 to 11, particularly in the rows corresponding to types **FRG**, **LOC**, **ZON**, **Town** and **Suburb**, represent the main core of our work. We can see that, overall, in GNP, only 34.9% of these features were matched (recall). While this seems a small number, it must be noted that the recall relative to YGP is much higher, 71.0%. Thus, the small number of pairs is due to the fact that GNP is more complete than YGP. Other reasons for this

**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 GeoPlanet$^{TM}$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 |

numbers include the errors associated with the geospatial coordinates of each ontology, or 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 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**). However, by querying Google Maps, we find that the location named "Serafão" closest to "Guimarães" is actually in "Fafe". This seems to suggest that the YGP relation is incorrect.

The partial results achieved after each step of the workflow figure in Table 12. This table represents only partial results of stage 2. For each section in the table, we show that the number of pairs decreases, which is a result of the fact that the work flow is in fact a series of filters that remove wrong or simply unsolved pairs.

## 4.1 Concordance RDF File

The most obvious choice for aligning the features in each ontology would be the *owl:sameAs* relation. However, this 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

**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 |
| | Total | 0.986 | 0.354 | 0.521 |
| YGP | 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 numbers corresponding to 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` | *17,702* |
| `manual pairs` | *15* |
| `semantic propagation` | 2,036 |
| `purge` | *1,131* |
| Total | *18,848* |

does not acknowledge a distinction between settlements and the respective civil parishes, also invalidates the use of that relation. Finally, aligning ontologies with the *owl:sameAs* relation can result in unwanted and unaccounted results, as shown in an example in SKOS documentation, [11], specifically in `http://www.w3.org/TR/2009/REC-skos-reference-20090818/#L4858`. For those reasons, we chose to use the *skos:closeMatch* relation instead, which is more appropriate and also reflects that the two concepts "can be used interchangeably in some information retrieval applications" [11].

On Postal codes, however, the relation must not be this one, since the entities aligned are not the same but related in a hierarchical manner. Thus, we used the *skos:broadMatch*.

### 4.2 SPARQL queries

Geo-Net-PT has a SPARQL [10] endpoint that can be used to query the ontology[1] [2]. We have included the alignment in the underlying database, allowing the queries to be made in GeoPlanet$^{TM}$ features:

```
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 .
}
```

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

**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) |

## 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$^{TM}$.

[1] `http://xldb.fc.ul.pt/wiki/Geo-Net-PT_02_SPARQL_endpoint`

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 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 system 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$^{TM}$ 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$^{TM}$.

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 the geospatial ontologies Geo-Net-PT and Yahoo! GeoPlanet$^{TM}$, using for the matching process information about the names and geospatial coordinates of the features. The results were stored in an RDF file that can be queried through a SPARQL endpoint or downloaded from the Geo-Net-PT web page [9]. The whole Geo-Net-PT ontology can also be downloaded from that page, as a PostgreSQL database dump file.

# References


[1] T. Berners-Lee, J. Hendler, O. Lassila, et al. 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.

[4] O. Lassila and R. R. Swick. Resource Description Framework (RDF) Model and Syntax, 1999.
`http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/`.

[5] V. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710, 1966.

[6] L. Lopes, F. Silva, F. Couto, J. Zamite, H. Ferreira, C. Sousa, and M. Silva. Epidemic Marketplace: An Information Management System for Epidemiological Data. *Information Technology in Bio-and Medical Informatics, ITBAM 2010*, pages 31–44, 2010.

[7] 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.

[8] M. Luczak-Rösch and R. Heese. Linked Data Authoring for Non-Experts. In *Linked Data on the Web Workshop*, 2009.

[9] XLDB Team, Faculty of Sciences, University of Lisbon. Geo-Net-PT, Accessed July 2010. `http://www.linguateca.pt/GeoNetPT/`, distributed by Linguateca.

[10] J. Perez, M. Arenas, and C. Gutierrez. Semantics and Complexity of SPARQL. *ACM Transactions on Database Systems (TODS)*, 34(3):16, 2009.

[11] W3C. SKOS Simple Knowledge Organization System Reference, 2009. `http://www.w3.org/TR/2009/REC-skos-reference-20090818/`.

[12] Yahoo. Yahoo! GeoPlanet™, Accessed July 2010. `http://developer.yahoo.com/geo/geoplanet`.