

# Mining a Multilingual Geographical Gazetteer from the Web

Adrian Popescu\*, Gregory Grefenstette\*\*, Houda Bouamor\*\*\*

\*TELECOM Bretagne (France), \*\*Exalead (France), \*\*\*LIMSI (France)

adrian.popescu@telecom-bretagne.eu, ggregens@exalead.com, houda.bouamor@limsi.fr

## Abstract

*Geographical gazetteers are necessary in a wide variety of applications. In the past, the construction of such gazetteers has been a tedious, manual process and only recently have the first attempts to automate the gazetteers creation been made. Here we describe our approach for mining accurate but large-scale multilingual geographic information by successively filtering information found in heterogeneous data sources (Flickr, Wikipedia, Panoramio, Web pages indexed by search engines). Statistically cross-checking information found in each site, we are able to identify new geographic objects, and to indicate, for each one, its name, its GPS coordinates, its encompassing regions (city, region, country), the language of the name, its popularity, and the type of the object (church, bridge, etc.). We evaluate our approach by comparing, wherever possible, our multilingual gazetteer to other known attempts at automatically building a geographic database and to Geonames, a manually built gazetteer.*

## 1. Introduction

Computer usable geographical information is needed for location-based services, for geographic information retrieval or for e-tourism platforms. In order for these applications to be truly useful, this information needs to be complete, accurate and applicable to the types of search requested. There exist manually built geographical gazetteers such as Alexandria [4] or Geonames [3] but their coverage is highly variable in different regions of the world and extending their coverage by hand would be a long arduous process. In addition, databases like Geonames do not contain popularity ranking information which would be useful in information retrieval applications. As an example of this limitation, it is currently impossible to select the main tourist attraction in France from Geonames. Recent works [8] [6], [11] describe ways to automate the constitution of large-scale geographical databases. [8] extracts names, GPS coordinates and a popularity measures for geographic objects from Flickr metadata

but no categorization of the objects nor spatial inclusion relations are extracted. In [6], we combined a statistical analysis and NLP techniques to extract place names, GPS coordinates, popularity measures and categorization information for geographic objects from Panoramio metadata. However, that analysis was limited to English, no spatial inclusion relations were extracted and the method was limited to Panoramio and to the English Wikipedia, exploiting thus only a small part of available georeferenced data sources.

In this paper, we describe a methodology to automatically create a multilingual geographical gazetteer by mining heterogeneous sources of georeferenced metadata (Flickr, Panoramio, Wikipedia) and AlltheWeb, a Web search engine. The method was applied to four major languages (English, French, Spanish and Italian) and it is easily adaptable to other languages.

## 2. Related work

[4] advises that each location in a gazetteer should be described by at least three elements: a name, GPS coordinates (or a spatial footprint) and a type. Geonames [3] and Alexandria [4] are two large manually constructed databases whose structures respect the minimal structure defined in [4]. Their elements are organized using two types of linguistic relations: *isA* (or *conceptual heritage*) and *partOf* (or *spatial inclusion*). Geonames [3] includes 8 main geographical type categories (such as *inhabited-locality*, *physical landmark* etc.), 645 intermediary ones (e.g.: *mountain* or *museum*) and more than 6 million place names (e.g. *Mont Blanc*, *Louvre*). Spatial inclusion (*partOf*) allows one to find that the *Louvre* is situated in *Paris*, which, is a part of *France*, part of *Europe* but these relations are not defined for all the elements of the database. In addition to these two types of relations, Geonames also contains precise localization coordinates, and alternative names of the object. Geonames was compiled from different existing databases and its coverage is highly variable from country to country (over 1.8 million entries for the

USA but only 24,000 for *Romania* or around 150,000 for *France*). Geonames covers *administrative divisions* and *hotels* well but often lacks information on other specific place names. For instance, for *Toulouse*, a large French city, there are less than 10 points of interest (other than *hotels*) listed in Geonames. Also, for a majority of the database entries, Geonames does not include any information which would allow ranking elements by importance or popularity, though one can see that such information would be useful in presenting retrieval results. In our system, we reuse the intermediary categories of Geonames to constitute a geographic vocabulary with which to discover new place names. We designed our resulting gazetteer to have a structure similar to Geonames, and since we capture complementary geographic information, our two resources can be merged easily.

Automating the acquisition of geographic information from Web sources has been proposed by [9] who introduces a notion of spatial context into text mining in order to extract information about landmarks. [11] also focuses on extracting vernacular names from Web pages. More closely related to our work, [8] was one of the first attempts to discover geographic names from a large volume of volunteered geographic information, exploring multiscale burst analysis to separate locations from others tags associated to georeferenced Flickr pictures. They report precision of 85% (with 50% recall) using a completely automatic analysis with no linguistic filtering of the resulting data. [1] exploited the results from [8] and associated a relevance value to each discovered place name. The resulting structure was used in an application for geographic image retrieval, with representative, popular tags overlaid on a scalable map. The database in [8] and [1] fails to respect the definition of a geographical gazetteer ([4]) because it does not contain categorization information, an important dimension of geographical gazetteers.

### 3. Gazetteer construction

Our work is based on detecting explicit geographic concepts in multiword place names (*Ponte* in *Ponte Vecchio*). A multilingual geographic vocabulary was constituted using Geonames and Wikipedia. We expanded the list of Geonames concepts available in English to French, Italian and Spanish using Wikipedia and then manually checked the validity of the translations. The obtained geographic vocabulary contains 300 concepts for each language. For instance bridge is represented as : *bridge, pont, ponte, puente*.

We identify georeferenced articles in Wikipedia (October 2008 dumps) using a list of 31 different geographic patterns, some common to all four

processed languages (“*coor dms*” or “*lat\_deg*”) and others specific to one language (“*latitudine:*” for Italian). After filtering the lists of georeferenced articles contained: 242 142 elements for English; 76 477 for French; 88 513 for Italian and 45 534 for Spanish. [2] performed a similar extraction reporting comparable results.

Since not all available Flickr and Panoramio georeferenced metadata are interesting for constituting a geographical gazetteer, we retained the following information: the textual description of the photo (text in Panoramio and tags and text in Flickr); the GPS coordinates of the images; the user ID and the photo ID. There is no explicit information concerning the language of the tags and texts and the same entity can be written in a variety of ways. For instance, the French version of the *Paris City Hall* appears as *Hôtel de Ville, Hotel de Ville* or *hotel de ville*. Given this variability, we preferred to transform all textual metadata into low-case ASCII text.

### 3.1. Algorithm for Flickr

Now we describe how place names are extracted, localized, attributed to encompassing regions, categorized and ranked by merging information from Flickr, Panoramio and AlltheWeb. Each geographic object extracted must respect the minimal definition of a place name given by Hill [4]. We add two supplementary dimensions, extracting also the encompassing entities and ranking the place names by importance, or popularity.

#### STEP1 - Place names detection

We apply the geographic vocabulary (1200 concepts) against the list of multiword Flickr tags and retain all tags that contain at least one geographic concept. Since we want to extract commonly used place names, we retain only multiwords that are employed by at least two different users of Flickr and Panoramio. Examples of place name candidates at this point include: *ponte vecchio, mia casa, crowded street, ppg place, tower bridge*. Separate lists are created for each language.

Some elements appear in more than one language because the respective geographic concepts are common to two or more languages. *mia casa* will be extracted for both Italian and Spanish because *casa* is common to the two languages; the same is true for *ppg place* because place is common to French and English. Language disambiguation is performed in STEP 5 of the algorithm. Place name candidates like *mia casa* or *crowded street*, which are not real place names are filtered out during the STEP 6 of the algorithm.

## STEP2 - Place name localization

Localization consists in averaging latitudes and longitudes for unambiguous candidate place names. Some place names are ambiguous, though. For instance, there is a *saint patrick's cathedral* in *Dublin*, one *New York* and a third *Karachi*. To separate the different instances of *saint patrick's cathedral*, we take the first occurrence of the name in the metadata and suppose that all the images inside a radius of 10 km that are tagged with this name depict the same object. We average the latitudes and the longitudes of all these photos and attribute the average values to the candidate place name. We consider that a different place named *saint patrick's cathedral* is discovered if a cluster of photos annotated with this multiword is found outside a 20 km radius from the GPS coordinates of the existing instance. This test is iterated over the entire Flickr and Panoramio metadata in order to discover all the different instances of an ambiguous place name. A name is retained only if its associated cluster of georeferenced metadata contains annotations from two different users and if the minimum distance between a two photos taken by two different users is smaller than a threshold. This step filters out some candidates like *crowded street* which are place names. For the experiments in Section 4, we empirically fixed the threshold value at 1500 meters.

Extended entities (*countries, regions*) are well covered in Geonames and our work focuses on precise entities (*buildings, parks*), which are less well covered. The 10 km radius was empirically determined after observing that a large majority of precise entities have a spatial extent smaller than the chosen radius.

## STEP3 - Encompassing entities finding

After STEP 2, each candidate name is characterized by a name and its (average) GPS coordinates. Now, we automatically mine a list of encompassing entities for each set of GPS coordinates. For instance, this means discovering that the coordinates of *musee d'orsay* (48.86, 2.32638) are situated in *Paris*, in *Île de France*, in *France*. Spatial footprints for *Paris*, the region *Île de France* and *France* are needed in order to perform this operation and, as we have mentioned, they are difficult to build automatically. Fortunately, georeferenced Flickr photos have associated spatial inclusion relations and it is simple to derive the list of encompassing entities for a (latitude, longitude) pair. To assign encompassing entities, we pick the georeferenced image closest to the GPS coordinates and use the Flickr supplied encompassing entities of that photo for the previously identified place name. Spatial inclusion relations in Flickr are also used in

order to create bounding boxes for extended place names (cities, regions, countries).

## STEP4 - Place name ranking

Ranking the elements of a gazetteer is useful in information retrieval applications in order to display the most relevant elements corresponding to a query [1]. In [6], we produced a ranking by aggregating information found in two complementary data sources, Panoramio and a search engine. Here, we use a slightly modified version of the ranking method first presented in [6]. The popularity rank of a place name is proportional to the number of different Flickr users that uploaded photos of tagged with that particular place name in 10 km radius around its GPS coordinates. By using spatial disambiguation, different ranks will be attributed to homonymous entities (i.e.: *saint patrick's cathedral* in *New York* and in *Dublin*). A significant number of place names will have an identical rank so we add a second component to the ranking measure. We launch an AlltheWeb query with the candidate place name plus the name of an encompassing entity (city if relevant; region name otherwise – for instance “**cathedral of learning**”+**Pittsburgh** or “**Mont Blanc**”+“**Haute Savoie**”). The addition of the encompassing region name is necessary to disambiguate the place name.

The ranking values are also useful for filtering out overspecified versions of well-known place names such as *eiffel tower*, since we have candidate place names in our list such as *blue eiffel tower* or *beautiful eiffel tower*. Note, that simple inclusion is not sufficient to eliminate these extra names, since not all terms that include a place name and something else are overspecifications (for instance, *one ppg plaza* and *ppg plaza*, respectively an office building and complex of buildings in Pittsburgh). We test each place name candidate against all other candidates within a radius of 10 km to detect possible overspecifications. If so, we compare the number of different users having uploaded photos tagged with the short and the long version of the candidate names and consider that an overspecification occurs if the ratio between the two values is superior to a threshold (which we set at 10).

## STEP5 - Place name categorization

The simplest way to categorize the geographical type of a place name would be to attribute it to the category that appears explicitly in the place name. This method, however, wrongly categorizes place names like *cathedral of learning* in Pittsburgh or *parc des princes* in Paris, which are not a *cathedral* and a *park*, but rather a *skyscraper* and a *stadium*. For terms that appear in several languages (*place kleber* is common to

French and English), we perform language identification prior to the categorization. We run successive AlltheWeb queries as in STEP 4 (place name + encompassing entity) for each place name, alternatively specifying “only English documents”, “only French documents” etc. and retain the most frequent language as pertinent for the place name.

Search engine results were already used for categorization tasks in [6] or [7]. In [6], we matched the geographic vocabulary against AlltheWeb results snippets corresponding to each place name candidate and retained the most frequent concepts in the snippets. Then, we formed definitional queries (“*candidate IS A category*”) with the three top ranked concepts in order to determine the category of the candidate. Definitional queries work well for well-known place names, which we categorized in [6], but often don’t exist for place names that are not that well-known. Here we only use the first phase, namely the extraction of the most frequent geographic categories from the snippets. We also realized that, when using snippets to categorize place names, many errors are generated by the co-occurrence of several place names in snippets. For instance *Champ de Mars* and *Tour Eiffel* co-occur frequently in the snippets associated to a query with “**Champs de Mars**+Paris. Co-occurring place names are usually written in capital letters and a simple way to diminish their negative influence on the categorization results is to consider only low-case terms in the snippets.

In [6], all candidates were assigned to one parent category. We discovered that errors appear mainly when the ratio of the counts associated to the top two categories in the snippets is small. We introduce a threshold in order to categorize only those place names with a ratio that is above this threshold.

### STEP6 - Place name filtering

Since place names are usually written in upper-case (*Tour Eiffel, Cathedral of Learning*) and other terms are usually written in lower case (*beautiful square, my house, red house*), we further exploit the content of snippets to filter out the unwanted terms. We count the total number of times a candidate appears in the snippets and the number of times when it is written in lower-case and eliminate it if the lower-case version is in majority. This filtering complements STEP2, where some non-place name candidates were eliminated based on distance constraints.

## 3.2. Algorithm for Wikipedia

Geographic data in Wikipedia have a richer structure compared to Flickr data and the algorithm for

extracting the same characteristics elements (place name, coordinates, type, encompassing entity and ranking) is in part different. Many of the different filters for Flickr data are not necessary in Wikipedia.

We consider the titles of georeferenced articles to be geographic names and their detection consists in simply isolating the title section in Wikipedia articles and, sometimes, eliminating some disambiguation information attached to the place names. For instance, we eliminate *Dublin* in *Saint Patrick’s Cathedral, Dublin* or *Nashville* in *Parthenon (Nashville)*. GPS coordinates are found using the same list of patrons we mentioned in the Preparatory Work section and encompassing entities are found in using the same method as for Flickr metadata (STEP3). The place names ranking is identical to the method described in STEP4.

In addition, we exploit Wikipedia links to find synonyms and to group together different translations of the same name. For instance (*Notre Dame de Paris, Cathédrale Notre Dame de Paris* and *Cattedrale di Notre Dame*) constitute a single entry. We also look for synonyms in the same language by checking if the articles contain an “*alternate\_name*” field and by extracting all the names this field includes.

The geographical type categorization procedure differs from the one described above in STEP5 because we can exploit the structure of Wikipedia articles. We observed that, when existing, the following Wikipedia articles parts are useful in categorization tasks: the Infobox, the first sentence; the article categories. We isolate these useful sections and match the geographic vocabulary against their content. As a result, we extract the parent category of the entry. We first check if there is an Infobox and, if so we extract the content of the “*Type*” field (*cathédrale pour Notre Dame de Paris*). Then we extract the geographic categories from the categories section: *Basilique, Cathédrale, Édifice*. Finally, we look in the first sentence in the article text for the first geographic concept (after the verb ‘to be’ or its translations), for example, “*Notre Dame de Paris (...) est la cathédrale de ...*”. After examining a large number of articles with Infoboxes, we observed that the type information was always correct and used it in priority. When there is no available Infobox, we combine information from the Categories section and from the first sentence and count the number of times each concepts appears in order to retain the most frequent. If there is a tie, then we favor geographical category types appearing in the first sentence.

At this point the processing, we have extracted the same place properties from both Flickr and Wikipedia.

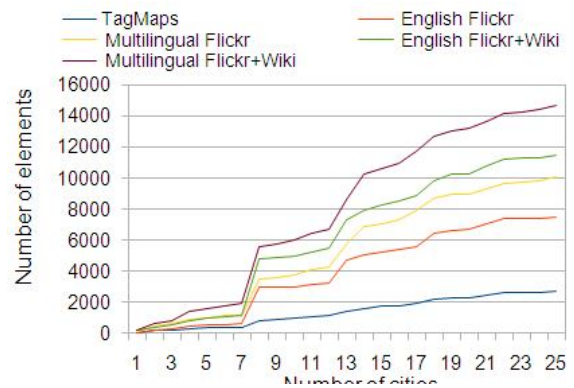
Each discovered place is characterized by its name, its GPS coordinates, its encompassing entities, its popularity, and its category. We compare the two lists of results and, when common elements are encountered, we favor Wikipedia results which come from a cleaner, better structured, edited source.

#### 4. Evaluation

The method for automatically constructing a multilingual gazetteer from user-contributed data, described here, can be compared to two other known attempts to automatically build such a database (TagMaps [8], Gazetiki [6]), as well as to Geonames, a manually created gazetteer and to TripAdvisor [10], an e-tourism platform. To do this, we manually selected 25 cities from different countries where different languages are spoken (examples include - English: *London, New York, Sydney*; French: *Montreal, Paris*; Spanish: *Mexico City, Barcelona, Valencia*; Italian: *Milan, Rome*; Others: *Beijing, Bucharest, Kiev, Tokyo*). We ran our place names selection algorithm and then evaluated the following: (i) the coverage of the multilingual Gazetiki against TagMaps and a English version of our database; (ii) the accuracy of the place name extraction process, with a focus on the performances of the filtering methods; (iii) the accuracy of the categorization procedure for place names extracted from Flickr, respectively from Wikipedia; (iv) the distance between the automatically attributed GPS coordinates and the Geonames coordinates for elements common to our database and Geonames; and (v) the accuracy of the ranking method by evaluated our ranking results and those in TagMaps against TripAdvisor, an external resource that contains ranked lists of tourist attractions. Excepting the localization procedure (iv), the other tests were performed manually because, although a somewhat tedious process, this was the only way to obtain accurate results.

##### 4.1. Database coverage

We selected a square area of around 125 km<sup>2</sup> around each evaluated city and compared the total number of place names in the multilingual version of Gazetiki to TagMaps [8] and to the English version of the database [6]. In figure 1, we present separate results for the use of Flickr as a data source and for a combination of Flickr and Wikipedia. All the results for Gazetiki correspond to a final version of the database obtained after all the different filtering processes. To obtain the TagMaps coverage for the 25 selected regions, we interrogated the TagMaps Web-service and recuperated all the tags in these regions.



**Figure 1. Coverage for multilingual Gazetiki, English Gazetiki and TagMaps. We present separate results for the use of Flickr and of Flickr and Wikipedia. Results are aggregated after the inclusion of each city.**

The coverage of all versions of Gazetiki (figure 1) is clearly superior to that of TagMaps, the lowest line in the graph. If we compare the results obtained when using only Flickr as a data source, the coverage of our database is three times greater than TagMaps for the English version (English Flickr) and four times higher for the multilingual version (Multilingual Flickr). When considering results for Flickr and Wikipedia (Multilingual Flickr+Wiki), the coverage of our database is nearly six times higher compared to TagMaps. After merging, 60% of our over 14,000 results came from Flickr and 40% came from Wikipedia. While we extract only multiword place names from Flickr, Wikipedia contains an important volume of georeferenced place names with no explicit geographic concepts in their name and the final coverage of our database is considerably increased.

The shapes of the curves in figure 1 show that the number of extracted place names varies from one city to another, depending on the number of georeferenced images in Flickr and Panoramio and on the description of the area in Wikipedia. Remarkably rich sets of results are obtained for London (3646) or Paris (1659), both three major cities that are well represented in Flickr, Panoramio and Wikipedia. A smaller number of results is obtained for Kiev or Toulouse (115). However, even in these cases, most important place names from these cities are represented in Gazetiki while this is not the case in TagMaps, where the corresponding coverages are Kiev (10) and Toulouse (20). The important coverage difference between Gazetiki and TagMaps is explained by the choice of a different place name selection method, combining statistical and linguistic information in our case and purely statistical for TagMaps. The constraints for

retaining a candidate place name are more relaxed in Gazetiki and the coverage improvement is doubled by higher quality of extracted elements.

We investigated the influence of multilinguism on the coverage of the gazetteer. Overall, the coverage of the multilingual version of Gazetiki increases by around 25% compared to the English corresponding English versions. For cities where local languages are French, Spanish or Italian, the number of place names in those languages is always greater than the number of place names in English. In Paris there are nearly 900 extractions in French and only 300 in English; in Rome there are over 300 extractions in Italian and fewer than 200 in English. The extension of Gazetiki to other languages would be beneficial for the coverage of the database and we plan on adding new languages to the framework.

As we mentioned, Geonames provides a good coverage of administrative divisions and hotels but a variable coverage of other place names. Excluding administrative divisions and hotels, there are 41 Geonames entries for *Athens*, none for *Bucharest*, there are four such entries for *Kiev* and 43 for *Mexico City*. Gazetiki contains 87 entries for *Athens*, 90 for *Bucharest*, 59 for *Kiev* and 134 for *Mexico City*. The coverage of the gazetteer is significantly improved compared to Geonames in all four cases. A qualitative analysis of Gazetiki shows that its entries are, in a majority of cases, tourist attractions. This is explained by the fact that Flickr photos related to places are usually taken and annotated by tourists.

#### 4.2. Place name extraction

We randomly selected up to 500 candidate place names (with a maximum of 20 per city) from our multilingual database and from TagMaps and counted the number of errors. We considered as correct all exact matches of the extracted instances to the real names (i.e. *Eiffel Tower* or *Squirrel Hill*) and incomplete matches which are commonly equivalent to their longer forms (i.e. *Notre Dame* instead of *Notre Dame Cathedral* or *Notre Dame de Paris*). A third category of results was constituted by generic names that can designate several specific entities (*church* or *catholic church*).

**Table 1. Precision of the place names extraction .**

	Only specific names	Including generic names
TagMaps	64.3%	74.7%
Gazetiki	96.2%	98.4%

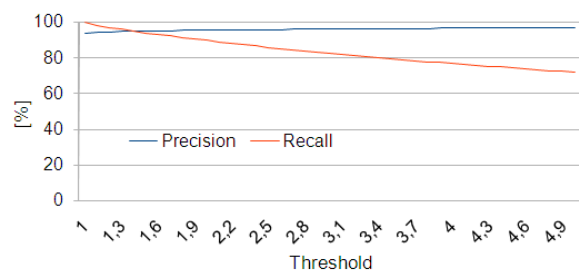
The results for Gazetiki (table 1) correspond to the final version of the database (after all the filters were applied) and they correspond to Flickr extractions.

The accuracy of the extraction is consistently better in Gazetiki compared to TagMaps in the two evaluation scenarios (96.2% versus 64.3% when we considering only specific names as correct and 98.4% versus 74.7% when including generic names). We consider that a geographical gazetteer should include only specific place names because generic names are more often than not associated to several entities, introducing noise in the database. However, if the primary purpose of an application is to increase the database, our method can be easily tuned in order to also extract generic names.

Since Flickr metadata are often noisy, we applied a series of filters in order to discard noisy place name candidates. We evaluated the efficiency of the filtering process on 500 place names that were initially extracted (STEP1) of the method but later discarded. Of the 500 eliminated candidates, 212 were in fact real place names and should not have been discarded. Among the correctly eliminated candidates, 113 were false place names (for instance *street photography* or *firenze piazza san firenze*); 73 were spelling errors (*ponte vechio* instead of *ponte vecchio*); 66 were overspecified terms (*eiffel tower paris france*) and 36 were generic names (*catholic church*). Here we focused on precision but if higher recall is needed, the filtering parameters can be easily tuned to obtain it.

#### 4.3. Place name categorization

The categorization procedure is different for place names extracted from Flickr (done exploiting Web search engine snippets) and from Wikipedia (done exploiting sections of the articles) and we present the results separately.



**Figure 2. Categorization performances considering the variation of the ratio between the frequencies of the most frequent geographic concepts in the.**

We used the same 500 candidates sample as for the extraction process in order to evaluate the

categorization of elements extracted from Flickr. Categorization accuracy was judged against a gold standard which was manually created. The influence of a threshold (the ratio between the frequency of the most frequent geographic concept appearing in the Web results associated to a place name and the second most frequent concepts) was studied and the results are reported in figure 2. When no threshold is applied (threshold = 1), the categorization precision is 93.2% and all elements in the evaluation set are categorized (recall = 1). Compared to [6], we count only low-case terms in the snippets when establishing the concept frequencies and do not use definitional queries anymore but the final results are improved compared to the ones reported in the cited paper (93.2% vs. 90%). The precision of the categorization improves when the value of the threshold increases but it is accompanied by an important recall reduction. When threshold = 5, the accuracy of the categorization is 96.3%, but only 71.9% of the candidates are categorized. The value of the threshold can be tuned in order to favor high precision or high recall, considering the applicative context in which the database is used.

In Wikipedia, we randomly selected 800 candidates (200 for each language) and manually verified the categorization results (table 2).

**Table 2. Categorization precision for ENGLISH, FRENCH, SPANISH, ITALIAN.**

	EN	FR	SP	IT
<b>Precision</b>	<b>94%</b>	<b>98%</b>	<b>95%</b>	<b>94%</b>

The categorization precision is good in all four languages, with a maximum of 98% for French and a minimum value of 94% for English and Italian. The use of different parts of the Wikipedia articles structure (Infobox, first sentence, categories) seems appropriate for categorization tasks and the results we present here are slightly improved compared to the use of the first sentence only [6].

In Flickr, errors appear mainly because some geographic concepts are well represented on the Web (*city, hotel* etc.) and they tend to have a high frequency in the Web results for any candidate place name. Wikipedia errors are notably caused by complicated first sentences and erroneous categories.

#### 4.4. Place name localization

The precision of the localization procedure was assessed only for place names extracted from Flickr because in Wikipedia the coordinates are directly provided in the structure of the article. We intersected the list of place names in Gazetiki to that in Geonames

and retained the common elements. Geonames GPS coordinates were considered a ground truth and we calculated the distance between these localizations and the GPS coordinates we extracted from Flickr tags.

A large majority of the automatically extracted GPS coordinates are close to the Geonames coordinates. 51.7% of the distances are inferior to 200 m and 82.3% are inferior to 1000 m whereas only 7.2% are superior to 3000 m. We looked at the distribution of distance differences with respect to the place name type and we observed that small distance differences are usually associated to entities that have a small spatial extent (*buildings*) whereas big distances are associated with place names with an important spatial extent (*islands, parks, rivers*). The differences between the GPS coordinates in our gazetteer and Geonames are mainly an effect of the fact that representative photos are often taken from a significant distance to the entity (especially for objects that have considerable spatial extents).

#### 4.5. Place name ranking

We evaluated the ranking of elements in Gazetiki and TagMaps by matching the top 10 place names in each city and each database to the 10 most popular tourist sites associated to each city in TripAdvisor. The ranking of elements in TripAdvisor is provided by the users and it can be considered to be a shared view of what is important/popular in each evaluated city. The capacity of Gazetiki and TagMaps to capture this shared view of cities summaries is evaluated by considering the common elements between the lists of top 10 ranked elements in the two databases and the TripAdvisor list. This intersection contains 62 elements for Gazetiki (out of 250) and only 15 elements for TagMaps and it shows that the popularity ranking in our database captures better city summaries compared to TagMaps.

In order to illustrate results, we present the top 5 place names associated to Barcelona and Moscow:

**-Barcelona:** Sagrada Familia; Montjuic; Tibidabo; Parc Guell; Casa Milla;

**-Moscow:** Red Square; Moscow Kremlin; Historical Museum; Novodevichy Convent; Moscow State University

The ranking procedure introduced in this paper generally succeeds in ranking best what seem to be the most representative place names for the analyzed cities.

## 5. Results discussion

We compared the multilingual gazetteer to two automatically built geographical databases and our results show significant improvements over these resources. Our place names extraction method based on the use of a geographic vocabulary returns results that are simultaneously more accurate and more detailed than the burst analysis implemented for TagMaps [8]. The criteria for retaining a candidate place name are more relaxed in our approach because we require that the extracted candidates contain a geographic concept, that it is used by at least two Flickr users and that the name is found in neighboring georeferenced photos. Candidates extracted from Flickr are often noisy but this series of filters eliminates most of the undesirable candidates (precision over 98% compared to 74% in TagMaps). Also, the popularity ranking procedure provides a better ranking than the one in [8] when compared to an external resource. Compared to [6], we modified the extraction procedure to adapt it to Flickr and extended the method to three new languages to the algorithm. The contribution of the newly added languages is particularly important for cities where the place names are unique to the local language.

We simplified the snippets based categorization procedure compared to the one in [6] and distinguished lower-case geographic concepts from others, a simple modification that has improved the quality of the results (over 93% compared to 90%). The categorization in Wikipedia exploits different parts of the articles whereas the procedure described in [6] only exploited the first sentence. This modification determines a slight improvement of the categorization precision (which exceeds 95% on average).

The localization procedure produces approximate GPS coordinates which are close to Geonames coordinates in a large majority of cases. The weak intersection between our gazetteer and Geonames (around 550 for over 14000 extracted entities) shows that the two resources are complementary. Since we use a domain model similar to, though richer than, that of Geonames, the integration of our new information into Geonames is straightforward.

## 6. Conclusion and future work

We introduced a method for automatically constructing a multilingual geographical gazetteer using heterogeneous data sources freely available on the Internet. Our gazetteer construction approach is mainly statistical and we intend to extend it to a large variety of languages. Here we discussed its application to four major languages but we are currently running

experiments with over 50 languages. More precisely, we plan to extend this work by (i) detecting place names based on term co-occurrences (extracting *Saint Sulpice and Paris* from photos tagged with *Paris, church, Saint Sulpice* etc.), (ii) introducing image reranking techniques to filter out irrelevant photos and (iii) detecting synonym place names (both intra- and inter-languages), (iv) detecting homepages for the place names, and maybe (v) visiting hours for tourist sites. Finally, we are exploring practical applications, such as geographic information retrieval or e-tourism platforms.

We acknowledge the need for multilingual geographic gazetteers and intend to make *Gazetiki* freely available by mid 2010.

## 7. Acknowledgements

This research is part of Georama, a French research project funded by ANR.

## 7. References

- [1] Ahern, S., Naaman, M., Nair, R. and Yang, J. 2007. World Explorer: Visualizing Aggregate Data from Unstructured Text in Geo-Referenced Collections. In *Proc. of JCDL 2007*.
- [2] Auer, S., Bizer, C., Lehmann, J., Kobilarov, G., Cyganiak, R. and Ives, Z. 2007. DBpedia: A Nucleus for a Web of Open Data. In *Proc. of ISWC 2007*.
- [3] Geonames - <http://geonames.org>
- [4] Hill, L. L., Frew, J. and Zheng, Q. 1999. Geographic names – the implementation of a gazetteer in a georeferenced digital library. *CNRID-Lib Magazine* (January, 1999).
- [5] Panoramio - <http://panoramio.com>
- [6] Popescu, A., Grefenstette, G., Moëllic, P.-A. *Gazetiki: automatic construction of a geographical gazetteer*. In *Proc. of JCDL 2008* (Pittsburgh, PA, June 2008).
- [7] Potrich, A. and Pianta, E.. L-ISA: Learning Domain Specific Isa-Relations from the Web. In: *Proceedings of LREC 2008*.
- [8] Rattenbury, T., Good, N., Naaman, M. 2007. Towards Automatic Extraction of Event and Place Semantics from Flickr Tags. In *Proc. of SIGIR 2007*.
- [9] Tezuka, T., Tanaka, K. Landmark Extraction: A Web Mining Approach. LNCS 3693. 379 – 396.
- [10] TripAdvisor – <http://tripadvisor.com>
- [11] Twaroch, F., Jones, C., Abdelmoty, A. Acquisition of a vernacular gazetteer from Web sources. In *Proc. of LocWeb Workshop 2008*.