Almeida, P. D., Rocha, J. G., Ballatore, A., & Zipf, A. (2016). Where the Streets Have Known Names. In International Conference on Computational Science and Its Applications. Berlin: Springer, pp. 1-12
10.1007/978-3-319-42089-9_1 [Author copy]

Where the streets have known names

Paulo Dias Almeida[1], Jorge Gustavo Rocha[1], Andrea Ballatore[2], and Alexander Zipf[3]

[1] Minho University, Portugal,
b6301|jgr@di.uminho.pt
[2] Birkbeck College, University of London,
a.ballatore@bbk.ac.uk
[3] Universität Heidelberg, Germany,
alexander.zipf@geog.uni-heidelberg.de

Abstract. Street names provide important insights into the local culture, history, and politics of places. Linked open data provide a wealth of knowledge that can be associated with street names, enabling novel ways to explore cultural geographies. This paper presents a three-fold contribution. We present (1) a technique to establish a correspondence between street names and the entities that they refer to. The method is based on Wikidata, a knowledge base derived from Wikipedia. a The accuracy of this mapping is evaluated on a sample of streets in Rome. As this approach reaches limited coverage, we propose to tap local knowledge with (2) a simple web platform. Users can select the best correspondence from the calculated ones or add another entity not discovered by the automated process. As a result, we design (3) an enriched OpenStreetMap web map where each street name can be explored in terms of the properties of its associated entity. Through several filters, this tool is a first step towards the interactive exploration of toponymy, showing how open data can reveal facets of the cultural texture that pervades places.

Keywords: Digital humanities, toponymy, OpenStreetMap, Wikidata, linked open data, volunteered geographic information

1 Introduction

All web maps show street names, supporting us in wayfinding. What is overlooked is that, behind each street name, that there is a rich and complex story. Street names are dedicated to notable people, places or events. They are frequently used to honor notable citizens or celebrate events and revolutions. Therefore, they often provide important insights into the culture, politics, and history of a locale.

In this pilot project we aim at creating an interactive web application where users can trace the stories behind street names, relying on OpenStreetMap[4] and other open data sources. As a first step, users can explore streets named after

[4] http://www.openstreetmap.org

individuals, filtering them by gender, date of birth, and profession. Wikipedia is used as an information source. More specifically, we use Wikidata[5] and DBpedia[6], two knowledge bases designed to extract structured information from Wikipedia, to link the street name with the corresponding resource described in the knowledge bases. To show the potential of linked open data, the process will be as automated as possible.

This paper describes the automatic mapping of street names with resources from these knowledge bases and rank those resources according to their relevance. The preliminary results, obtained on a sample of streets in Rome, show that there are many missing relations. To increase the coverage, we propose a web tool to that knowledge from human contributors.

The remainder of this paper is organized as follows. We start by presenting related work in Section 2. We then elaborate on our approach in Section 3. Section 4 evaluates our automated solution, and the preliminary results are discussed in Section 5. The design proposal for the web platform that expands and complements the automated solution is presented in Section 6. Finally, we present our conclusions in Section 7.

2 Related work

To link street names to the relevant entities, we adopt concepts and techniques from a variety of research areas, including toponymy, geographic information science (GISc), and Semantic Web and Linked Open Data research.

2.1 Toponymy and street names

Toponymy is the study of place names (toponyms), with respect to their origins, meanings, use and typology. Place names provide an extremely useful geographical reference system in the world. Consistency and accuracy are essential in referring to a place to prevent confusion in everyday activities. Toponymy is crucial to establish officially recognized geographical names, and relies local written and oral histories to study and record how place names evolve and why.

Many geographers, historians, and linguists have found that toponyms provide valuable insight into the historical geography of a particular region. They play a symbolic role in the expression of local culture, being used many times to promote values related to political and religious beliefs [7]. Unsurprisingly, place names are then given an important role in territorial conflicts and landscape transformation [9]. Place names are so important that, even outside of armed conflicts, altering place names in official maps to reflect a different context and culture is regarded as a possible act of cultural aggression [9]. Consequently, place names represent an extremely important data source for analysing cultural changes across different locations over time.

[5] `http://www.wikidata.org`

[6] `http://wiki.dbpedia.org`

2.2 Linked open data

Linked data aims to provide knowledge in a structured and simple manner, allowing it to be understood by humans and machines [2]. This is done through the adoption of design principles and standards in order to express data in the simplest form. Data is organized in a set of triples; subject, object, and a named relation connecting them (predicate). This design goal of establishing named and directed links between typed data enables the creation of useful semantic queries and facilitates the integration of heterogeneous data sources.

The linked data paradigm has also emerged as a promising approach to structuring and sharing geospatial information. The simplification in the way data is expressed has a dramatic impact in spatially and temporally referenced data, usually modeled as complex relational schemata [10]. User-generated content, of major importance for many current applications, can also benefit from the linked data approach. Triples are seen as statements made by a know author, with great potential applications in collaborative geographic information production.

The process of linking a new dataset to existing ones is called 'bootstrapping', and is usually performed on semantic hubs, such as Wikidata and DBpedia [12, 4]. These are important community efforts to extract structured information from Wikipedia according with linked data principles. These projects allow you to ask sophisticated queries against datasets derived from Wikipedia and to link other datasets on the Web to Wikipedia data.

The process of inter-linking data is complex and faces several technical and cultural challenges. Knowledge bases are built on heterogeneous and incompatible vocabularies and ontologies [3]. Several efforts have been made to ease and automate the linking process, using semantic similarity and relatedness measures [1].

3 Street name matching method

Our main goal is to establish links between street names and the entities that the street names refer to. This mapping will then be used to enable a different visualization of streets and neighbourhoods, putting these entities and their historical and cultural context. We start by retrieving the street names from OpenStreetMap. Each toponym is then used in a semantic query to a knowledge bases such as Wikidata and DBpedia to retrieve relevant entities. As a case study, we selected street names in Rome, expressed in Italian.

3.1 Toponym retrieval

In order to identify the entity represented by a street name, we start by isolating the part of the name that corresponds to said entity. Street names are normally composed with prefixes or suffixes that describe the feature type (examples in English include *avenue*, *street*, and *boulevard*). These linguistic tokens have to be filtered out before querying the knowledge base. For this purpose, we use a set of stop words. This process is language-specific and entails the a priori definition of the stop words used as prefixes and suffixes for each target language–Italian in this case study.

3.2 Entity retrieval

After having filtered the street name, the resulting string is used in a query to the knowledge bases. In order to be incorporated in the process, a knowledge base needs to support a text-based query for relevant entities, and their properties and relations. In order to make the process as automated as possible, all the query endpoints and parameters, as well as result format, can be defined in a configuration file. This way, the application can interact with any knowledge base that supports either SPARQL or HTTP queries. The pilot tool works seamlessly with Wikidata and DBpedia, two major open knowledge bases.

3.3 Entity ranking

After receiving the results from a knowledge base, we have to estimate their relevance. Each time we query the knowledge base for a keyword, several entities can be retrieved. Some street names in Rome, like *Via Mazzarino* will match multiple resources in Wikipedia. Mazzarino can be either a place name[7] or the surname of several notable people.[8] In order to establish which entity is referred to by a street name, a ranking of the relevance of the results needs to be calculated. Working with semantically rich data results in more adequate rankings than simple keyword-based search. Properties like the entity's location provide additional context to the query to determine the relevance of a result.

Our ranking algorithm takes into account the location of the input street and of the entities. This approach has already been proven to generate more appropriate rankings in related projects [6, 11], and is crucial to this application, where the local context plays a major role. Based on the process defined by Shuyao et al. [11], we define a formula to incorporate location into a custom ranking algorithm. The relevance r of an entity e is defined as follows:

$$r_e = \beta \; i_e + (1-\beta)(1 - dist(\lambda_q, \lambda_e)) \tag{1}$$

where i_e represents the informativeness associated with the entity, λ_q and λ_e represent the locations of the query and of the entity respectively, and β is a factor between 0 and 1 used to balance the importance of the informativeness calculated for the entity in relation to the distance between the query and an entity.

In order to calculate the informativeness of an entity (i_e), we first estimate its global relevance. In order to calculate this relevance we chose to consider the number of different language entries Wikipedia has for said entity. However, this

[7] `https://it.wikipedia.org/wiki/Mazzarino_(Italia)`
[8] `https://it.wikipedia.org/wiki/Mazzarino`

approach is biased in favor of very general entities, such as countries, cities, and family names. To solve this problem, we weight the informativeness of an entity with its inverse entity frequency (ief), similar to the method by Zaragoza et al. [13]:

$$ief = \log(N/n_e) \tag{2}$$

where N represents the totality of query results and n_e those that contain entity e. In the next section, the results obtained by this entity ranking algorithm are explored.

4 Evaluation

To validate the methodology presented and assess the suitability of the custom ranking algorithm, we proceed to evaluate the performance of our approach. For this evaluation, a central area of Rome, including the neighborhoods of Trastevere and Testaccio, was chosen as the target geographic area. The set of named streets used was obtained with the following query, on the OpenStreetMap API called Overpass:[9]

```
[out:json];
(
  way
    ["highway" = "residential"]["name"]
    (41.8642593, 12.4612841, 41.9030756, 12.4945021)
);
out body;
>;
out skel qt;
```

The query originated a set of 1,709 named streets. When applied to this set, our street name matching algorithm found an entity for 66% of the streets. To evaluate the quality of the matching, 20 queries with more than one entity were randomly selected. 79 entities were mapped to these queries, with the number of entities associated per query varying from two to seven. Then, the relevant entities for each query were manually selected and ranked from most to least relevant. The set of relevant documents per query is thus limited, in this evaluation, to those returned by the knowledge base for the selected sample. Table 1 illustrates an example of the process executed on Wikidata.

For this sample, the proposed solution was able to map a relevant entity for 85% of the queries, however, only about 35% of the retrieved entities were considered relevant results. This fact accentuates the importance of ranking the relevance of these results. Hence, we take as base line the default ranking returned by the knowledge base, and proceed to analyse how it compares with the custom ranking. The performance of the ranking algorithms is based on

[9] http://overpass-turbo.eu

Table 1. Sample query used in the evaluation. *Sommergibile* means submarine.

Street name	"Via Galileo Ferraris"
Query	"Galileo Ferraris"
Wikidata entities	1. Galileo Ferraris (sommergibile 1914)
(default ranking)	2. Galileo Ferraris (sommergibile 1935)
	3. Galileo Ferraris (Italian physicist)
Wikidata entities	1. Galileo Ferraris (Italian physicist)
(custom ranking)	2. Galileo Ferraris (sommergibile 1935)
	3. Galileo Ferraris (sommergibile 1914)
Top entity	Galileo Ferraris (Italian physicist)

Table 2. Performance of the ranking methods, using mean average precision (mAP), mean reciprocal rank (MRR), and normalized discounted cumulative gain (nDCG). Best results in bold.

Algorithm	mAP	MRR	nDCG
Default	0.70	0.74	0.74
Custom	**0.75**	**0.76**	**0.80**

the mean average precision (mAP), mean reciprocal rank (MRR), and normalized discounted cumulative gain (nDCG), three common information retrieval measures. Table 2 summarizes the comparison of these two ranking approaches.

Our custom ranking outperforms the baseline in all three measures, thanks to the inclusion of location awareness and entity informativeness, calculated from the number of different Wikipedia language entries. The custom ranking algorithm proposed outperforms the default ranking, demonstrating a 7% relative increase in mAP and a 8% relative increase in nDCG.

Despite demonstrating encouraging results, this evaluation also brings to light a significant level of noise, in the form of incorrect entities being mapped to the queries. This can be a significant problem, regardless of the quality of the ranking algorithm, suggesting a coverage issue in the entities in the knowledge base. The limited scope of the evaluation sample is also a point of concern, highlighting the necessity to involve local communities in the matching and validation process.

Through a crowdsourced solution, users will be able to either select the best correspondence from the suggested ones, or add new entities when the automated process fails. The custom ranking algorithm developed for the automated part of the solution can also benefit from this user interaction, learning better disambiguation strategy. Cases where the user solves a conflict between entities of different classes, for example, a street is matched to an entity representing a person and also to a number of entities representing places named after that person, can be learned by the ranking algorithm to apply in the subsequent queries.

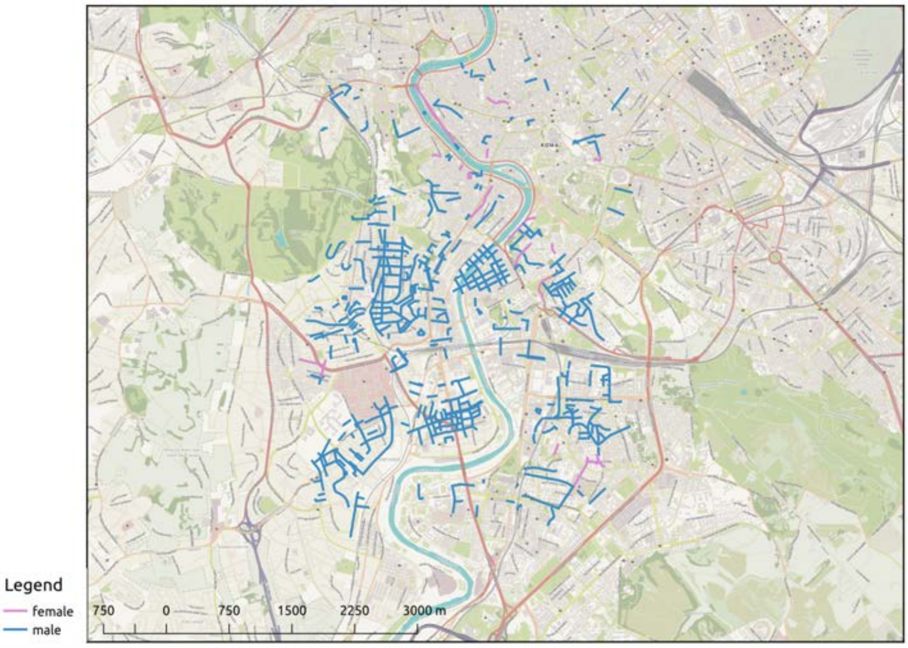

Fig. 1. Gender of people after whom the streets in Rome are named.

5 Results and discussion

As stated in the previous section, our solution was tested on an area of Rome, and for the 1,709 retrieved roads with names, of which 1,121 were matched to Wikidata entities. This association between street names and known entities in a knowledge base allows for analysis over the properties and characteristics of the entities referred to by street names.

First, we consider the streets that are named after people. In this case, we determined that 630 of the 1,121 matched streets are named after individuals (approximately 56%). Subsequently, we observe the gender of the referred people. As is possible to notice in Figure 1, only approximately 6% (36 instances) are named after females, while a staggering 94% of names refer to males.

Based on the information contained in Wikidata, we can also analyze the thematic areas in which these individuals became notable. Figure 2 shows that, for this region of Rome, the top three thematic areas of activity are science (109 instances), religion (90 instances), and politics (84 instances). The map also attracts attention to the existence of spatial clusters of streets named after writers, scientists, and politicians, reflecting clear planning choices of the Roman administration.

This analysis exemplifies the richness and depth of the cultural data contained in Wikidata and similar knowledge bases, and how, following the linked data paradigm, cultural exploration of the geographic space can be supported.

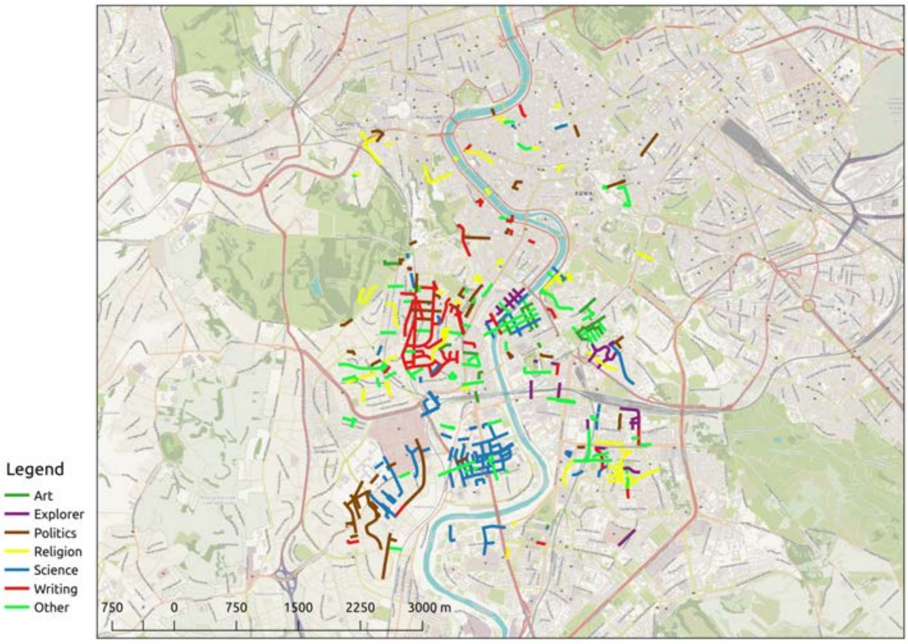

Fig. 2. Professions of people after whom the streets in Rome are named.

However, as mentioned in Section 4, there are some important limitations. First, there is a significant number of irrelevant entities being mapped to street names, introducing some noise in the data. Second, the coverage of the knowledge base is particularly limited for entities that are notable only in local contexts, such as people whose existence never became known beyond their hometown. The next section proposes a solution to these issues.

6 Crowd sourcing local knowledge

The previous section discussed preliminary results obtained with our street name matching approach. Both the OpenStreetMap streets and the Wikidata knowledge base used in this study were built through crowd sourcing. It is therefore reasonable to get the communities involved, taking advantage of the local knowledge of citizens [8]. The locals, in fact, are the ones that are most likely to interpret the meaning of their local street names. Hence, these volunteer communities can contribute to the linking process as follows:

- From the lists of related entities, they can pick the best match based on their locally-situated knowledge.
- If only irrelevant entities are listed, they can add a link to Wikipedia or other web resources that refer to the relevant entities.

To achieve this goal, we outline a dedicated web platform. Designing a successful crowd sourcing platform is not easy. In another words, the success of such platforms does not depend only on technical aspects, such as design principles and the usability of tools [5], but are deeply rooted in the motivation of contributors. To advance the design of this platform, we outline its requirements, present a story board documenting a typical use case, and describe its current software architecture.

6.1 Requirements

The requirements for this VGI platform are derived directly from the aforementioned goals:

- The user will be able to navigate a map and pick any street;
- The entities associated to a selected street are presented in a list;
- The user has the power to remove/add entities from this list;
- The user can select any entity from the list as the best match for that street;
- The user can contribute with properties/relations missing from an entity;
- The user can access a page with revision history for the match and propose a change;
- The user can access a history of his contributions;
- Users can see a listing of overall top contributors to the platform;
- The user will be able to contribute with a Wikipedia login.

6.2 Storyboard

According to the requirements, the web platform will have as its main screen a map where the user can select any street and visualize its associated entities. The user also has access to a discussion page were the review history can be accessed. The sketch in Figure 3 represents this scenario. In the case where a suitable match still has not been selected for the street, the user will be able to see the list of associated entities and edit as desired. The user can also select any entity in the list as the best match for the street. This part of the interface is represented by the sketch in Figure 4.

6.3 Software architecture

In terms of its overall architecture, this application will try to integrate and take advantage of other collaborative projects. The raster tiles for the map and the vector layer with street information is provided by OpenStreetMap. Tiles can be retrieved from several OSM tile servers and vector data can be efficiently obtained using the Overpass API.

Wikipedia will serve as the source for the information associated with each entity, complemented by its structured knowledge bases (Wikidata and DBpedia). The mapping between streets and entities, along with associated revision history, will be stored in a PostgreSQL/PostGIS database.

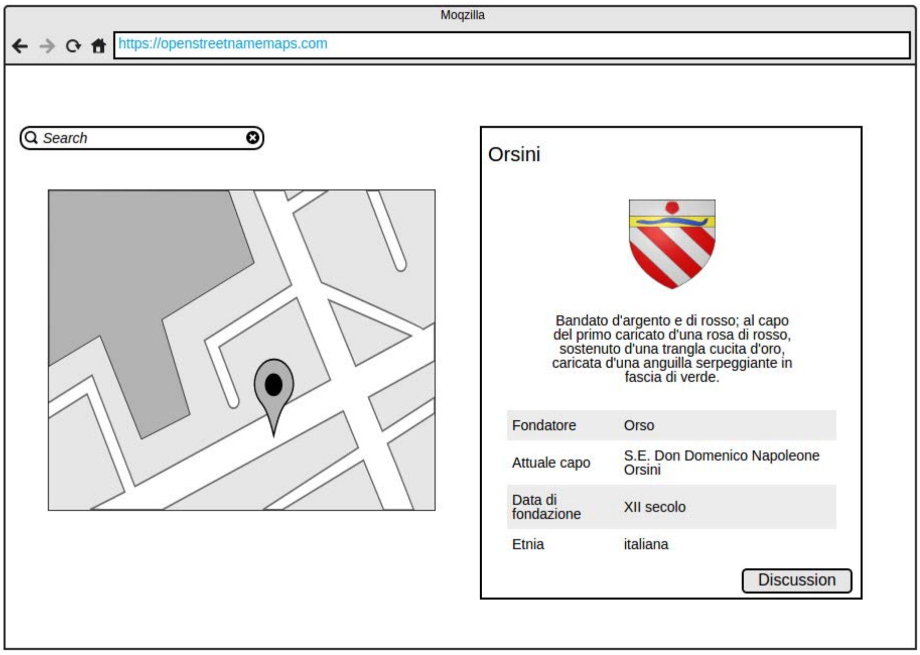

Fig. 3. Selecting a street on the web platform.

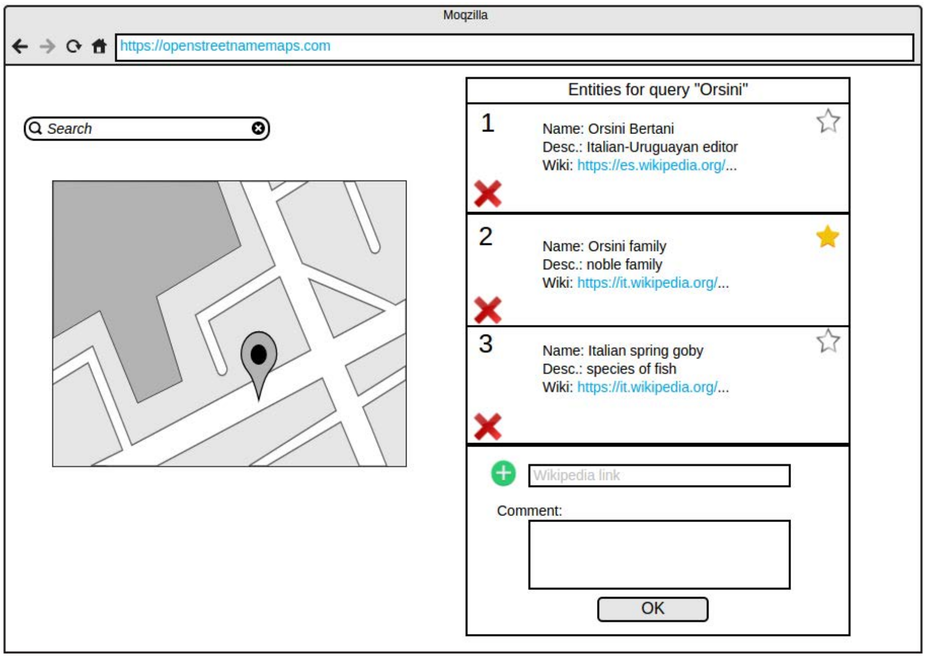

Fig. 4. Editing list of associated entities on the web platform.

The authentication can be provided by Wikipedia OAuth service API. Users can use their existing credentials. The platform must follow the best practices that has been proven to reliably lead to community engagement and participation. For this reason, the application must keep track of who did what and when, providing the full history of changes, allowing quick undo for error and vandalism correction. All these software components will be arranged in an open source tool, freely usable and editable.

7 Conclusion and future work

In this paper we presented an approach to create and improve the mapping between OpenStreetMap street names and entities represented in structured knowledge bases. The ultimate goal is to generate maps where each street name can be explored in terms of the properties of its related entity, enabling a deeper analysis that provides insights into the geographic culture, history, and politics of a place. The automated mapping of entities with street names was complemented by the implementation of a custom ranking algorithm, which improved upon the default ranking obtained from the knowledge base. After identifying the limitations of an automated solution, we designed a web platform to better fulfill our goal.

This approach will enrich the semantics of OpenStreetMap and Wikipedia, creating a new, machine-readable information layer that connects toponyms with known entities. This knowledge can be returned to the community to be freely used for general exploration, tourism, and research. In the short term, we will make this mapping available online to be easily processed with other tools and linked to other knowledge bases. In parallel, we will continue the development of the web platform and its algorithms, extending it from the Italian case to other languages. Increasing access to this kind of cultural geo-information will trigger new and engaging ways to navigate and understand the intricacies of places.

References

[1] Ballatore, A., Bertolotto, M., Wilson, D.: Linking geographic vocabularies through WordNet. Annals of GIS 20(2), 73–84 (2014)

[2] Ballatore, A., Wilson, D., Bertolotto, M.: A Survey of Volunteered Open Geo-Knowledge Bases in the Semantic Web. In: Pasi, G., Bordogna, G., Jain, L. (eds.) Quality Issues in the Management of Web Information, Intelligent Systems Reference Library, vol. 50, pp. 93–120 (2013)

[3] Ballatore, A., Mooney, P.: Conceptualising the geographic world: the dimensions of negotiation in crowdsourced cartography. International Journal of Geographical Information Science 29(12), 2310–2327 (2015)

[4] Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: DBpedia-A crystallization point for the Web of Data. Web Semantics: Science, Services and Agents on the World Wide Web 7(3), 154–165 (2009)

[5] Brown, M., Sharples, S., Harding, J., Parker, C., Bearman, N., Maguire, M., Forrest, D., Haklay, M., Jackson, M.: Usability of geographic information: Current challenges and future directions. Applied Ergonomics 44(6), 855–865 (2013)

[6] Buscaldi, D., Magnini, B.: Grounding Toponyms in an Italian Local News Corpus. In: Proceedings of the 6th Workshop on Geographic Information Retrieval. pp. 15:1–15:5. GIR '10, ACM, New York (2010)

[7] Cohen, S.B., Kliot, N.: Place-names in Israel's ideological struggle over the administered territories. Annals of the Association of American Geographers 82(4), 653–680 (1992)

[8] Goodchild, M.F.: Citizens as sensors: the world of volunteered geography. GeoJournal 69(4), 211–221 (2007)

[9] Kadmon, N.: Toponymy and geopolitics: The political use–and misuse–of geographical names. The Cartographic Journal 41(2), 85–87 (2004)

[10] Kuhn, W., Kauppinen, T., Janowicz, K.: Linked Data: A Paradigm Shift for Geographic Information Science. In: Duckham, M., Pebesma, E., Stewart, K., Frank, A.U. (eds.) Geographic Information Science, LNCS, vol. 8728, pp. 173–186. Springer, Berlin (2014)

[11] Qi, S., Wu, D., Mamoulis, N.: Location aware keyword query suggestion based on document proximity. IEEE Transactions on Knowledge and Data Engineering 28(1), 82–97 (2016)

[12] Vrandečić, D., Krötzsch, M.: Wikidata: A Free Collaborative Knowledge Base. Communications of the ACM 57(10), 78–85 (2014)

[13] Zaragoza, H., Rode, H., Mika, P., Atserias, J., Ciaramita, M., Attardi, G.: Ranking Very Many Typed Entities on Wikipedia. In: Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management. pp. 1015–1018. CIKM '07, ACM, New York (2007)