# MATCHING ALTERNATIVE ADDRESSES: A SEMANTIC WEB APPROACH

S. Ariannamazi [a], F. Karimipour [b,*], F. Hakimpour [b]

[a] Department of Geographic Information Systems, Faculty of Civil Engineering, Kerman Graduate University of Technology
[b] Faculty of Surveying and Geospatial Engineering, College of Engineering, University of Tehran, Tehran, Iran
(sobhan.namazi, fkarimipr, fhakimpour)@ut.ac.ir

**KEYWORDS:** Semantic web, Semantic integration, Literal matching, Address matching

## ABSTRACT:

Rapid development of crowd-sourcing or volunteered geographic information (VGI) provides opportunities for authoritatives that deal with geospatial information. Heterogeneity of multiple data sources and inconsistency of data types is a key characteristics of VGI datasets. The expansion of cities resulted in the growing number of POIs in the OpenStreetMap, a well-known VGI source, which causes the datasets to outdate in short periods of time. These changes made to spatial and aspatial attributes of features such as names and addresses might cause confusion or ambiguity in the processes that require feature's literal information like addressing and geocoding. VGI sources neither will conform specific vocabularies nor will remain in a specific schema for a long period of time. As a result, the integration of VGI sources is crucial and inevitable in order to avoid duplication and the waste of resources. Information integration can be used to match features and qualify different annotation alternatives for disambiguation. This study enhances the search capabilities of geospatial tools with applications able to understand user terminology to pursuit an efficient way for finding desired results. Semantic web is a capable tool for developing technologies that deal with lexical and numerical calculations and estimations. There are a vast amount of literal-spatial data representing the capability of linguistic information in knowledge modeling, but these resources need to be harmonized based on Semantic Web standards. The process of making addresses homogenous generates a helpful tool based on spatial data integration and lexical annotation matching and disambiguating.

## 1. INTRODUCTION

The rapid developments of cities and changes in aspatial attributes of spatial features has made it difficult to find new addresses which indicates the necessity of a system to match addresses more than ever. On the other hand, Earth and space science researches and applications often collect and analyze a large amount of geospatial data. However, most of the geoscience information are not obtained directly from measurements but rather derived from other data by the application of a scientific workflow in which each analytical step consumes and produces data with particular representations. In recent years, scientific workflows are emerging as a suitable practice to model and simulate the logical stages of a science process to create a science product (Ludascher et al., 2006). A visionary concept of the integration of geo-information was posed on 1998 by the U.S. vice president Al Gore. His "Digital Earth" label became popular for describing a virtual representation Of the Earth on the Internet that is spatially referenced and interconnected with the world's digital knowledge archives (Vacari et al., 2009). However, rapid development of crowd-sourcing or volunteered geographic information (VGI) provides opportunities for authoritative that deal with geospatial information. Allowing amateurs to collect geospatial data helps lower the cost, capture richer user-based information and reflect real world changes more quickly. At the same time it may also dilute information quality, such as completeness, consistency and accuracy (Jackson et al. 2010). Heterogeneity of multiple data sources and inconsistency of data types is an intrinsic characteristics of VGI data because almost everyone using a VGI database can change the information. This policy is considered as a strength since it allows collecting more thorough data with less resource needs, and also a weakness from the aspect of redundancy and inconsistency of the dataset. The redundant and inconsistent data has to be managed in a manner that users be able to rely on and trust the quality of the datasets. The crowd-source information has potential to be the best source of free information without any requirement for technical sensors, however, one of the most frequent challenges one might encounter when using such sources is the lack of reliability.

Semantic web is a potent technology developed to achieve the web of data. However, it has more capabilities such as finding alternatives in different sets of vocabularies referring to the same entities which makes it perfect for the matching process we are approaching. Although, this is still a work in progress we were able to find the alternatives with good accuracy. On the other hand, due to the variations in the feature attributes the processing is more efficient in a graph database (triple store) than a relational database. This, once again indicates the capabilities of semantic web technology as an integrating tool for the datasets.

Since geospatial ontologies for authoritative and volunteered data sets are developed independently, matching geospatial ontologies is an essential step to use them synergistically. Ontology matching is the task of finding a mapping, i.e. a set of correspondences, between entities from different ontologies. It includes two main levels, the terminology level and instance level. Many ontology matching methods and systems have been developed in recent years (Shvaiko and Euzenat 2012). In geospatial information science, several data conflation methods have been developed for matching or integrating geospatial vector data, mainly based on the similarities of geometries or topological relations, as well as attributes, if available. Most of them focus on conflating road vector data.

Ontology refers to an explicit specification of a shared conceptualization and plays an important role in establishing shared formal vocabularies. A spatial individual has a certain and verifiable location and a meaningful label, which together distinguish itself from others (H. Du et al 2013). To be effective, geo-spatial applications need to provide powerful and flexible search capabilities to support their users. However, discovery services are often limited by only syntactically matching user terminology to metadata describing geographical resources (Shvaiko et al. 2010). Since geospatial ontologies for authoritative and volunteered data sets are developed independently, matching geospatial ontologies is an essential step to use them synergistically.

Compared to other ontologies, geospatial ontologies have some special properties. Firstly, many geospatial terminologies are commonly used in daily life and their meanings vary in different contexts. For example, ‘‘College’’ may refer to an institution within a university in one ontology, whilst meaning a secondary school in another. In addition, geospatial ontologies often do not have a huge number of classes as ontologies in several other subject areas (for example, biomedicine) do, but may represent many real world spatial individuals, whose locations, at least in theory, can be verified (H. Du et al 2013).

The Semantic Web is an extension of the Web through standards by the World Wide Web Consortium (W3C). The standards promote common data formats and exchange protocols on the Web, most fundamentally the Resource Description Framework (RDF). According to the W3C, "The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries". Below are some of the most frequently used standards of semantic web:

RDF is a general method to decompose any type of knowledge into small pieces, with some rules about the semant.css, or meaning, of those pieces. The point is to have a method so simple that it can express any fact, and yet structured enough that computer applications can do useful things with it.

The OWL 2 Web Ontology Language, informally OWL 2, is an ontology language for the Semantic Web that became a W3C Recommendation on Oct 27 2009. OWL 2 ontology documents describe information in terms of classes, properties, individuals, and data values the relationships of which can be described by a number of features.

SPARQL is an RDF query language, that is, a semantic query language for databases, able to retrieve and manipulate data stored in Resource Description Framework (RDF) format.

A URI is simply a Web identifier: like the strings starting with "http:" or "ftp:" that you often find on the World Wide Web. Anyone can create a URI, and the ownership of them is clearly delegated, so they form an ideal base technology with which to build a global Web on top of. In fact, the World Wide Web is such a thing: anything that has a URI is considered to be "on the Web".

## 2. METHODOLOGY

The expansion of cities resulted in the growing number of POIs in the Open Street Map, the most well-known VGI source, that cause the datasets to outdate in short periods of time so these changes made to spatial and aspatial attributes of features such as names and addresses might cause confusion or ambiguity in the processes that require feature’s literal information like addressing and geocoding. VGI sources neither will conform specific vocabularies nor will remain in a specific schema for a long period of time. As a result, the integration of VGI sources is crucial and inevitable in order to avoid duplication and the waste of resources.

Information integration can be used to match features and qualify different annotation alternatives for disambiguation. Unfortunately, the changes in aspatial attributes of the features, like name, do not conform a specific discipline and in result the integration and matching process cannot be achieved using NLP (Natural Language Processing) methods. However, the spatial attributes of these features will remain tolerably the same. As a matter of fact, these spatial attributes are the main reason such features would be considered the same by common people. A spatial individual has a certain and verifiable location and a meaningful label, which together distinguish itself from others. Geospatial ontologies describe conceptual hierarchies and interrelations of terminologies in the domain of geo-spatial science, which are used to describe facts (classifications, relations, attributions and locations) about spatial individuals.

Having seamless and homogenous datasets, addressing and geocoding procedures can be accomplished more efficiently and according to the minimal client’s knowledge -including programs and users- about the region. Consequently, finding matches in such identical attributes could be one of the best solutions used in the integration process. GeoSPARQL is a standard for representation and querying of geospatial linked data for the Semantic Web from the Open Geospatial Consortium (OGC). The definition of a small ontology based on well-understood OGC standards is intended to provide a standardized exchange basis for geospatial RDF data which can support both qualitative and quantitative spatial reasoning and querying with the SPARQL database query language. GeoSPARQL implements the topology relation discovery which can be used to achieve the required matching process.

There are (almost) no complete implementations of GeoSPARQL at the moment, there are, however partial or vendor implementations of GeoSPARQL. One of the most complete implementations of GeoSPARQL is used by Parliament Triple Store an open source graph database frequently used for semantic web purposes. To interact with Parliament we have developed an application using C# language and .NetRDF a library for working with RDF data based on “.NET Framework” platform. The application currently supports the simple style of “Main Street –

Secondary Street – Alley”. The process consists of finding the alternative names for given feature names based on topology relations and producing an alternative address using the results.

Fig 1. demonstrates the overall procedure of reaching a consistent and homogenous dataset. There are two kinds of matching processes used in the process: Ontology matching and literal matching. In the recent years several methods for matching ontologies have been presented. On the contrary matching the literals has almost been neglected.

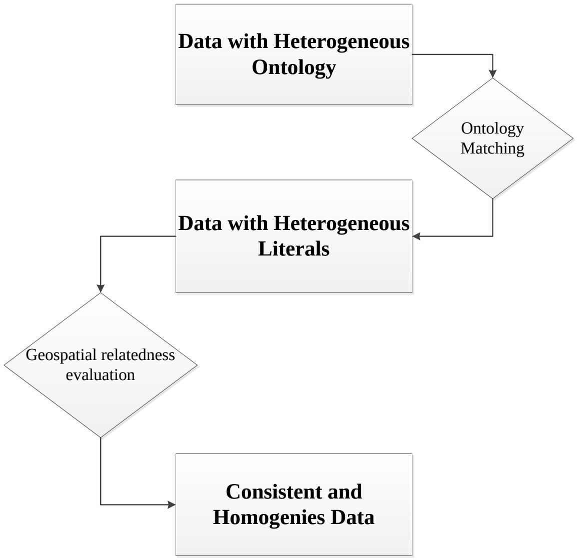

Data with Heterogeneous Ontology
Ontology Matching
Data with Heterogeneous Literals
Geospatial relatedness evaluation
Consistent and Homogenies Data

Figure 1. Processing procedures for matching

## 3. RESULTS AND DISCUSSION

In this research we have used a sample datasets of former and current names of the features for the city of Kerman (Figure 2). One of the challenges we were facing during this research was the lack RDF data for Kerman therefor we had to create such data. In order to convert available datasets which is in relational tables or XML formats in most cases, datalift software which is an application able to convert many file formats to RDF is utilized. However, the shp2rdf (a java library for converting shape files to RDF formats) still has some bugs so we were forced to change the RDF schema to GeoSPARQL using .net RDF and C# (Figure 3). Converted data contained 22863 nodes connected to each other by 13848 ways.

As mentioned in the introduction, we consider the geometry attribute of alternative features to remain nearly the same. On the other hand, due to the difference between the datasets in question we cannot consider them equal, thence, we have to use a set of buffers with 5 to 30 meter distance around the features to overcome this issue. Although, this method improves the matching results, there are still some exceptions which cause the process not to be completely accurate (i.e. some streets or alleys being too close to each other, or, streets being expanded over time).

So making a decision about the semantic similarity between two features depends on the amount of the vicinity and relatedness of the features in question. The approximation of the features are validated between two data sets in question including OSM and a relational database derived from the local agency (municipality).

Further semantic similarity assessment improve the probability of matching procedure. So the process of semantic similarity matching is based on the matching of the instances and not the concepts or schemas of the two databases in hand. We have considered the similarity as the inverse of the distance between the objects. They take on large values for similar objects and either zero or a negative value for very dissimilar objects. A key aspect of this implementation is that it is possible to obtain a vector that is derived based on the number of the words in the phrases in question. Passing the aforementioned two-step procedure and by comparing and weighting geometrical relatedness and semantic similarity, it is more comfortable to decide whether two feature are the same or not.

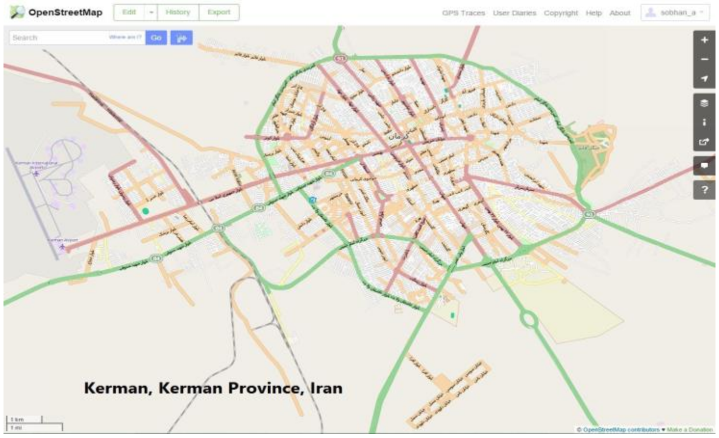
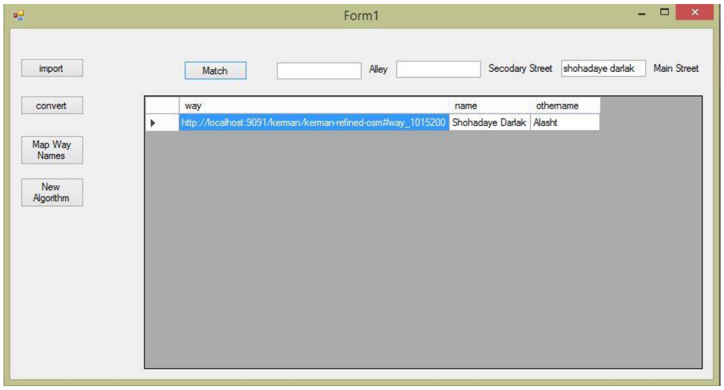


Figure 2. The Study Area



shohadaye darlak

| | way | name | othemame |
|---|---|---|---|
| ▸ | http://localhost:9091/kerman/kerman-refined-osm#way_1015200 | Shohadaye Darlak | Alasht |

Figure 3. Developed Matching Application Using C#

## 4. CONCLUSION

In the present study two distinct datasets are compared and their geospatial equality. The developed software enables whom that may concern to have more accurate information about the changes and differences in the names of the streets. Using the results of this comparisons a specific address.
Using developed software one can compare an address or a set of addresses to the address annotations of a reference dataset to estimate whether a specific address falls within an address domain corresponding to a feature in the reference dataset. If an address falls within a feature's address domain, and after the approval of geometrical similarity approval it can be considered a match and a location can be returned. Future study include providing different addresses corresponding to different paths from a place to a destination.

## REFERENCES

Ludascher, B., Altintas, I., Berkley, C., Higgins, D., Jaeger, E., Jones, M., Lee,E., Tao, J., Zhao, Y., 2006. Scientific workflow management and the Kepler system. Concurrency and Computation: Practice and Experience 18, 1039–1065.

Du, Heshan, Natasha Alechina, Mike Jackson and Glen Hart. "Matching Formal and Informal Geospatial Ontologies." In *Geographic Information Science at the Heart of Europe*, edited by Danny Vandenbroucke, Bénédicte Bucher and Joep Crompvoets, 155-171: Springer International Publishing, 2013.

Vaccari, Lorenzino, Pavel Shvaiko and Maurizio Marchese. "A Geo-Service Semantic Integration in Spatial Data Infrastructures." *International Journal of Spatial Data Infrastructures Research* 4, (2009): 24-51.

Jackson MJ, Rahemtulla H, Morley J (2010) The synergistic use of authenticated and crowdsourced data for emergency response. The 2nd international workshop on validation of geoinformation products for crisis management (VALgEO).

Shvaiko P, Euzenat J (2012) Ontology matching: state of the art and future challenges. IEEE Transactions on Knowledge and Data Engineering.

Shvaiko P, Ivanyukovich A, Vaccari L, Maltese V, Farazi F (2010) A semantic geo-catalogue implementation for a regional SDI. INPSIRE Conf.

Li, M., Zhang, Y., Zhu, M., & Zhou, M. (2006, July). Exploring distributional similarity based models for query spelling correction. In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics (pp. 1025-1032). Association for Computational Linguistics.