Grounding Linked Open Data in WordNet: The Case of the OSM Semantic Network*

Andrea Ballatore,[1] Michela Bertolotto,[1] and David C. Wilson[2]

[1] School of Computer Science and Informatics
University College Dublin, Ireland
{andrea.ballatore,michela.bertolotto}@ucd.ie
[2] Department of Software and Information Systems
University of North Carolina, Charlotte, NC
davils@uncc.edu

Abstract. In recent years, the linked open data (LOD) paradigm has emerged as a promising approach to structuring, publishing, and sharing data online, using Semantic Web standards. From a geospatial perspective, one of the key challenges consists of bridging the gap between the vast amount of crowdsourced, semi-structured or unstructured geo-information and the Semantic Web. Notably, OpenStreetMap (OSM) has gathered billions of objects from its contributors in a spatial folksonomy. The contribution of this paper is twofold. First, we add a piece to the LOD jigsaw, the OSM Semantic Network, structuring it as a W3C Simple Knowledge Organization System (SKOS) vocabulary, and discussing its role in the constellation of geo-knowledge bases. Second, we devise *Voc2WordNet*, a mapping approach between a given vocabulary and WordNet, a pivotal component in the LOD cloud. Our approach is evaluated on the OSM Semantic Network against a human-generated alignment, obtaining high precision and recall.

Keywords: Geo-semantics, OpenStreetMap, Linked Open Data, OSM Semantic Network, WordNet, Semantic alignment, Semantic mapping, Voc2WordNet.

1 Introduction

Since its invention in the early 1990s, the World Wide Web (WWW) has enabled an unprecedented growth of digital data, offering a platform for publishing, retrieving, and sharing any type of data across the globe. An enormous volume of data has been disseminated online in a variety of formats, resulting in an archipelago of incompatible data spaces. A crucial limitation to the full exploitation of this ocean of heterogenous data is the lack of clear semantics, which hinders the ability to analyse, explore, and discover unexpected connections and relations between entities.

* The research presented in this paper was funded by a Strategic Research Cluster grant (07/SRC/I1168) by Science Foundation Ireland under the National Development Plan. The authors gratefully acknowledge this support.

S. Liang, X. Wang, and C. Claramunt (Eds.): W2GIS 2013, LNCS 7820, pp. 1–15, 2013.

A prominent attempt to overcome this structural limitation of the WWW, and provide a unified platform for data semantics, is Berners-Lee's Semantic Web [10]. One of the most successful outcomes of this ambitious initiative is the so-called linked open data (LOD) paradigm, with the purpose of creating a unified data space. To be classified as LOD, data must be released under open licenses; saved in a machine-readable digital format; stored in non-proprietary formats; accessible via URIs; and linked to other LOD [9].

As LOD is generated and published online, a growing web of inter-linked datasets has emerged, resulting in the LOD cloud, also referred to as the Web of Data, defined by Bizer et al. [11] as "a web of things in the world, described by data on the Web" (p. 2). The more linked data is available, the more connections can be discovered between datasets, exploiting network effects to deliver rich and relevant results to users [21]. Large linked data repositories are maintained online.[1] Recently, the commercial potential of the paradigm has been highlighted by Google's Knowledge Graph, a large semantic artifact that utilises Freebase, an LOD resource, to semantically enrich the search engine's results [30].

As a large part of online data involves a spatial component, geographic open data is a first class citizen in the LOD cloud [6]. Semantics is key to enabling the usage, integration, and exploration of geographic data [1, 21]. The advantages of the LOD paradigm applied to geographic information are particularly evident in the context of geographic information retrieval (GIR), where existing techniques have shown limited effectiveness [28]. A linked data search engine such as DBpedia Faceted Search promises – and often returns – highly relevant results to complex geospatial queries, such as 'Rivers that flow into the Rhine and are longer than 50 kilometers.'[2]

The emergence of the LOD infrastructure has a great potential for the dissemination of geographic data. A prominent example is found in the British Ordnance Survey, which has embraced the paradigm and released some of its resources as linked data [19]. In parallel, volunteered geographic information (VGI) is gaining credibility as a source of detailed information generated by non-expert users through crowdsourcing [13]. Challenging traditional top-down cartographic engineering, OpenStreetMap (OSM) provides an open platform to build a world map, tapping its contributors' knowledge of their local geographic context [12]. To date, a gap between VGI datasets and the LOD cloud exists, and constitutes a barrier to the integration and usage of the data.

In this paper, we contribute to bridging the gap between VGI and the LOD cloud in two ways. First, we describe how we have structured the OSM Semantic Network using the W3C Simple Knowledge Organization System (SKOS), and published online as LOD. The OSM Semantic Network offers a machine-readable, structured, open conceptualisation of OSM semantics, and constitutes a semantic support tool to interpret, search, and tap the project's vast vector dataset. We originally extracted the network from the OSM Wiki website and other sources to compute the semantic similarity of geographic terms [5]. Second, we outline and

[1] See for example `http://thedatahub.org` (acc. Oct 30, 2012).
[2] `http://wiki.dbpedia.org/FacetedSearch` (acc. Oct 30, 2012).

evaluate *Voc2WordNet*, a semantic mapping technique to connect OSM terms to WordNet synsets, enabling the discovery of rich semantic relations between terms such as part-whole (e.g. part-of relations) and subsumption (e.g. is-a relations). This semantic mapping is not a goal in itself, but can enable a number of search operations on both OSM and WordNet.

The remainder of this article is organised as follows. Section 2 reviews relevant work in the areas of LOD, open geo-knowledge bases, OSM semantics, semantic mapping, and WordNet. Section 3 presents an LOD resource extracted from OSM semantics, the OSM Semantic Network. Section 4 describes and formalises a generic approach to semantic mapping onto WordNet. Subsequently, we report on the evaluation of the approach, executed on a subset of terms from the OSM Semantic Network (Section 5). This paper concludes with a summary of results and directions for future research in Section 6.

2 Related Work

OSM has received wide attention, generating a large number of academic studies and commercial projects. This section surveys related work relevant to the OSM Semantic Network, VGI, and WordNet, with respect to geo-semantics and the LOD paradigm.

2.1 OpenStreetMap Semantics

From its foundation in 2004, OSM has established itself as the most ambitious VGI project [12]. From a semantic viewpoint, OSM is a semi-structured folksonomy, which allows contributors to create any new term to describe the objects that they find worth mapping [32]. This radically open approach to geo-semantics is supported by the fact that an all-encompassing geographical ontology is an unrealistic endeavour, and that a bottom-up negotiation allows for more experimentation, and attracts non-expert contributors. As project founder Steve Coast [12] succinctly put it, "to dictate [terms] as in a top-down ontology would have been nuts." The downsides of the adoption of a semi-structured folksonomy include wide variability and ambiguity in the interpretation of terms, proliferation of near-synonym terms, and lack of explicit semantic relations, resulting in a 'spatially rich and semantically poor' dataset [4].

In recent years, efforts have been undertaken to strengthen the thin semantic ground on which OSM rests, including LinkedGeoData [2], and OSMOnto.[3] Baglatzi et al. [3] devised an approach to grounding the OSM folksonomy on the DOLCE upper-level ontology [17]. Acknowledging the extreme difficulty in implementing such semantic mapping in an automatic way, they designed a game with a purpose (GWAP) to crowdsource a human-quality mapping. In our previous work, we devised an initial semantic integration between OSM and DBpedia, geared towards exploratory navigation of Web maps [4].

[3] `http://wiki.openstreetmap.org/wiki/OSMonto` (acc. Oct 30, 2012).

To tap the knowledge contained in the OSM Wiki website, we extracted the OSM Semantic Network via a dedicated open source crawler. An early, off-line version of the semantic network was utilised to compute the semantic similarity of OSM terms using link-based measures [5]. In this paper, we extend the OSM Semantic Network by re-structuring it as a SKOS vocabulary, integrating it in the LOD cloud, and devising a mapping technique to WordNet.

2.2 WordNet as Semantic Ground

Since the early 1990s, WordNet has been a precious semantic resource for many applications in natural language processing and artificial intelligence [16]. The core element of WordNet is the 'synset,' a concept that represents set of synonymous words, called 'word senses.' WordNet has found particular success in the areas of word sense disambiguation and semantic similarity [26, 7]. Different components of the network have been exploited to model the semantic similarity of its synsets, tapping its deep taxonomy, and the word definitions, called 'glosses' [e.g. 29]. Although the semantic network was not designed for this purpose, it has been frequently used as an upper level ontology, i.e. a general-purpose semantic ground, for example to discover semantic connections in unstructured data [22].

From a geospatial viewpoint, GeoWordNet aggregates WordNet synsets with the open gazetteer GeoNames [18]. To date, none of the numerous alternative semantic resources has yet managed to dethrone WordNet from its leading position as a general-purpose semantic ground. In the context of the LOD cloud, WordNet is used as a high-quality primary information source in many projects [6]. The lexical database is a well-established linked dataset, wired to a number of open knowledge bases.[4] These resources are inter-linked with DBpedia, a core node of the LOD cloud. In this paper, we devise a general technique to map a vocabulary onto WordNet, using it as a limited, and yet rich semantic ground.

2.3 Open Data Integration

To generate LOD, it is necessary to link the new entities to existing ones in the LOD cloud, a process often called 'bootstrapping' [23]. The identification of the same concepts and entites in heterogenous data spaces is crucial to supporting the Semantic Web. Merging different conceptual schemas is a time-honoured challenge in computer science, started well before the advent of the WWW. Logical reasoning, machine learning, and statistical analysis have been utilised to tackle the problem in the context of database schemas [27].

The Ontology Alignment Evaluation Initiative (OAEI) has proposed benchmarks and performance metrics specifically tailored to the area of ontology alignment and integration [15]. Several approaches to generating a mapping have been devised, both from an intensional and an extensional viewpoint. *Terminological* methods rely on simple string matching between the terms, while *semantic* methods compare the representation of terms in formal semantic models. Furthermore, *internal* methods observe aspects of the terms in isolation, such as

[4] `http://www.w3.org/2006/03/wn/wn20/` (acc. Oct 30, 2012).

Table 1. Namespaces of the OSM Semantic Network and related datasets

Abbr.	Description	URI
osn	OSM Semantic Network	`http://spatial.ucd.ie/lod/osn/`
owl	OWL	`http://www.w3.org/2002/07/owl\#`
rdfs	RDF schema	`http://www.w3.org/2000/01/rdf-schema\#`
dc	Dublin Core	`http://purl.org/dc/elements/1.1/`
skos	SKOS	`http://www.w3.org/2004/02/skos/core\#`
wn	WordNet synset	`http://www.w3.org/2006/03/wn/wn20/instances/synset-`
ws	– word sense	`http://www.w3.org/2006/03/wn/wn20/instances/wordsense-`
wns	– schema	`http://www.w3.org/2006/03/wn/wn20/schema/`
lgdo	LinkedGeoData	`http://linkedgeodata.org/ontology/`

the attribute ranges. By contrast, *external* methods analyse the relational structure of the ontologies, comparing the position of the terms relative to the other terms. Finally, *extensional* methods perform the alignment based on distributional properties of term instances.

Despite the variety of existing mapping techniques, to the best of our knowledge, a semantic mapping technique between a vocabulary and WordNet, geared towards the 'bootstrapping' of the vocabulary in the LOD cloud, has not been devised. *Voc2WordNet* has the purpose of filling this specific gap. As described in Section 4, *Voc2WordNet* performs the semantic mapping between a vocabulary term and a specific WordNet word sense from an intensional (i.e. lexical overlap between the lexical definitions) and an extensional perspective (i.e. the usage frequency). The next section describes our contribution to the area of VGI in the LOD cloud.

3 The OSM Semantic Network as Open Data

The OSM Semantic Network is a semantic artifact containing the conceptualisation of OSM tags, which we developed in our previous work to provide a semantic support tool for OSM.[5] The artifact can be used to compute the semantic similarity of tags [5]. In this section, we report on how the OSM Semantic Network has been structured using W3C Simple Knowledge Organization System (SKOS), and published online in the LOD cloud.

From a semantic viewpoint, OSM is a semi-structured folksonomy. The terms are documented on the OSM Wiki website, in an open process of semantic negotiation and consensus-building. Unsurprisingly, the consistency in the actual usage and intended meaning of these terms is rather low, resulting in semantic ambiguity that hinders the possibility of exploiting the project's rich vector dataset [25]. The OSM Semantic Network provides a machine-readable structure that can support the automatic manipulation of OSM features in data mining, GIR, and information integration.

Initially developed as an offline dataset, the OSM Semantic Network has been integrated in the LOD cloud. In order to facilitate the exploration and usage of

[5] `http://wiki.openstreetmap.org/wiki/OSMSemanticNetwork` (acc. Oct 30, 2012).

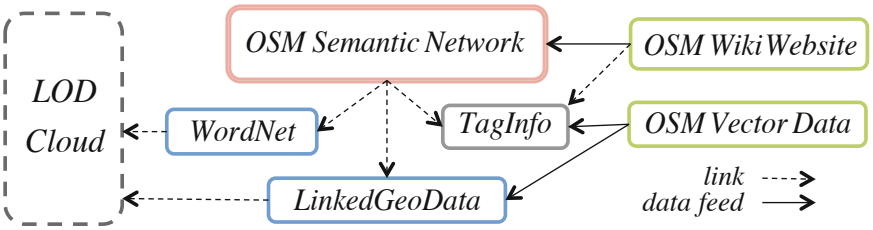

Fig. 1. The OSM Semantic Network in context

the network, we have published it online with a human-readable web interface.[6] Figure 1 shows the location of the OSM Semantic Network in the context of LOD, and the data flow from and towards related projects, including OSM, LinkedGeoData, WordNet, and TagInfo. For the sake of brevity, all the URIs in the remainder of this article are shortened (see Table 1).

We have structured the OSM Semantic Network as a SKOS vocabulary [24]. SKOS is a semantic formal language designed to allow the publication and sharing of technical vocabularies, taxonomies, and classification systems. In a SKOS scheme, the main semantic unit is the *skos:Concept*. A concept is a term that can be defined using lexical definitions and linked to other concepts through semantic relations. The semantic relations are explicitly left as generic as possible. Concepts can be more general or specific than other concepts (*skos:broader* and *skos:narrower*), and can be semantically related (*skos:related*).

Hence, each term defined in the OSM Wiki website corresponds to a SKOS concept. As the URIs are a key asset in LOD, the mapping between OSM tags and OSM Semantic Network terms is direct and intuitive. For example, the tag *waterway=river* corresponds to the term *osn:term/k:waterway/v:river*. The quality of the SKOS vocabulary was assessed based on the criteria outlined by Suominen and Hyvönen [31]. The OSM Semantic Network is linked to the LinkedGeoData ontology, via about 660 *skos:exactMatch* relations. Our approach to grounding a given vocabulary in WordNet is described in the next section.

4 *Voc2WordNet*, a Semantic Mapping Algorithm

This section presents *Voc2WordNet*, an algorithm devised to generate a semantic mapping between a vocabulary and the lexical database WordNet. The algorithm generates a semantic mapping between a given vocabulary V containing a set of terms (e.g. a SKOS vocabulary), and WordNet synsets that are semantically similar. *Voc2WordNet* can be used to map any vocabulary onto WordNet, enabling some degree of interoperability. More formally, a semantic mapping m between term $t \in V$ and synset $s \in W$ with relation r has the form of a triple $< t, r, s >$. In the OSM Semantic Network, we define a fine-grained semantic

[6] Pubby, available at `http://www4.wiwiss.fu-berlin.de/pubby` (acc. Oct 30, 2012).

mapping, based on the SKOS mapping relations.[7] Hence, *Voc2WordNet* generates three symmetric mapping relations:

Exact (*skos:exactMatch*): Identical terms that can be used interchangeably with high confidence (e.g. 'university' in OSM and LinkedGeoData). This relation is logically equivalent to *owl:sameAs*.

Close (*skos:closeMatch*): Similar terms that might contain some contradiction, and therefore cannot engage in identity (e.g. 'wood' in OSM and 'forest' in WordNet).

Related (*skos:relatedMatch*): Terms that are semantically related by a non-hierarchical relation (e.g. 'power station' in OSM and 'electricity' in WordNet). This relation is non-transitive.

The purpose of *Voc2WordNet* is to obtain correct mappings $m =< t, r, s >$ between the vocabulary V and the WordNet synsets W. For example, the definition of *wn:gallery-noun-3* is "a room or series of rooms where works of art are exhibited." By contrast, *wn:gallery-noun-1* is defined as "spectators at a golf or tennis match," and *wn:art-noun-1* as "the products of human creativity; works of art collectively." Hence, the desired mappings are *<osn:Art_gallery close wn:gallery-noun-3>* and *<osn:Art_gallery related wn:art-noun-1>*.

Voc2WordNet generates a set M of mappings m between a given vocabulary V and the set of WordNet synsets W. Given a term $t \in V$, *Voc2WordNet* utilises a lexical matching function on the words contained in t, taking compound words into account (e.g. 'swimming pool'), and then splitting them if not defined in WordNet (e.g. 'swimming' and 'pool'). If the set of matching wordsenses ws is not empty, the algorithm relies on three indicators of semantic salience:

Word sense frequency f: The usage frequency f of a WordNet word sense is correlated with its semantic salience. In the context of a shared vocabulary, common word senses are more likely to be correct than uncommon word senses. For example, for $t =$'field', *ws:field-noun-1* ("a piece of land cleared of trees and usually enclosed") has a usage frequency $f = 49$, whilst *ws:field-noun-12* ("all of the horses in a particular horse race") has $f = 1$. Indeed, this assumption can be false in the context of open text.

Lexical overlap *ol*: Similar terms tend to be defined using the same words. The lexical overlap *ol* is the number of word shared by two terms. Terms showing high lexical overlap are more likely to be salient than terms that do not show overlap. The overlap is considered after the removal of stopwords, and lemmatisation, excluding the term that is being defined. For example, the overlap between the definitions of term t ("A river is a body of water") and *wn:river-noun-1* ("Rivers are natural streams of water") is equal to 1.

Salient taxonomy Θ: If a vocabulary is domain specific, the mapping can be restricted to a salient taxonomy Θ, i.e. a subset of WordNet. Salient word senses tend to engage in semantic relations with salient synsets. Looking at the noun taxonomy of WordNet, it is possible to select high-level synsets that

[7] `http://www.w3.org/TR/skos-reference/\#mapping` (acc. Oct 30, 2012).

are salient to the vocabulary's domain. If the candidate synsets engage in some relation with such salient taxonomical roots, they are more likely to be valid than synsets that do not. For example, let us choose *wn:artifact-noun-1* as a salient root, and 'shelter' as t. It is possible to infer that *ws:shelter-noun-2* ("protective covering that provides protection from the weather") is related to the salient root through a path of transitive subsumption relations (*wns:hyponymOf*), while *ws:shelter-noun-4* ("a way of organizing business to reduce the taxes it must pay on current earnings") is not.

Formally, we define t as the input term, C_t as the set of candidates for term t, ws as the candidate word sense, s as the corresponding synset, and Θ as a manually selected salient taxonomy. The non-negative θ is set to 1 if $s \in \Theta$, and 0 otherwise. The salience of the three indicators are captured in a normalised score σ as follows:

$$\sigma(t, ws, s) = \frac{2|C_t| - rank(f(ws)) - rank(ol(t, s)) + \theta}{2|C_t| - 1} \tag{1}$$
$$\sigma \in [0, 1],\ rank \in [1, |C_t|]$$
$$\theta = 1\ if (s \in \Theta), \theta = 0\ otherwise$$

The salience score σ captures the semantic similarity between term t and the synset s, through the word sense ws, relative to the set of candidates C_t. The ranking function $rank$ is applied on the set C_t, and returns an integer between 1 and $|C_t|$. The score falls in the interval $[0, 1]$, where 0 indicates no salience, and 1 maximum salience. For example, given a C_t with three candidates, if ws and s have the highest frequency ($rank(f) = 1$), the second highest overlap ($rank(ol) = 2$), and s belongs to the salient taxonomy Θ ($\theta = 1$), then $\sigma = .8$.

In order to provide more flexibility, the algorithm filters out candidates based on a minimum frequency (f_{min}), and a minimum overlap (ol_{min}). Once the candidate having the highest σ has been selected, an appropriate relation r must be chosen from the set { *exact*, *close*, *related* }. As a selection heuristic, we define three boolean conditions, i.e. $rank(f) = 1$, $rank(ol) = 1$, and $s \in \Theta$. If all of the three conditions are true, $r = exact$; if at least two conditions are true, $r = close$; othwerwise $r = related$. The detailed workings of the algorithm are outlined in Algorithm 1. In the next section, *Voc2WordNet* is evaluated on a real-world scenario, i.e. a subset of the OSM Semantic Network.

5 Evaluation

This section describes a preliminary experimental evaluation of *Voc2WordNet*, applying the semantic mapping technique to the OSM Semantic Network. The technique obtains a high-precision mapping between the terms defined by the OSM Semantic Network and WordNet. First, we generate an evaluation dataset M_h (Section 5.1). Second, we define performance measures (precision and recall) that compare the machine-generated mapping M with the human mapping

Algorithm 1. *Voc2WordNet*$(V, W, ol_{min}, f_{min}, \Theta)$

input : vocabulary V, set of synsets W, min overlap ol_{min}, min word sense frequency f_{min}, salient taxonomy Θ
output: Set M of semantic mappings $m < \ldots t, r, s >$

```
M ← ∅
foreach term t ∈ V do
    m ← findSemanticMapping(t, W_t);
    add m to M;
    extract terms from lexical definition of t to set D_t;
    foreach term d ∈ D_t do
        m_d ← findSemanticMapping(d, W_t)
        add m_d to M;
return M.
```

Function findSemanticMapping(t, W_t)

```
C_t ← ∅
foreach ws ∈ W_t do
    find set of matching word senses ws ∈ W_t with lexicalMatch;
    find synset s corresponding to ws in WordNet;
    fetch word sense frequency f(ws) from WordNet;
    compute lexical overlap between definitions ol(s, t);
    apply filters f_min and ol_min;
    compute salience score σ(s, ws, t);
    add pair < s, ws > to candidate set C_t;
select best candidate s_b ∈ C_t having max(σ(s, ws, t));
select relation r ∈ { exact, close, related };
generate mapping m =< t, r, s_b > and return it.
```

M_h (Section 5.2). An experiment on a number of parameter combinations is executed (Section 5.3), and the performance of *Voc2WordNet* is discussed and summarised (Section 5.4).

5.1 Ground Truth

To construct a mapping gold standard, we select a random sample of 30 terms from the OSM Semantic Network, corresponding to the 0.6% of the entire dataset. The sample terms were manually mapped to semantically salient WordNet synsets. By manually selecting correct mappings between the 30 terms from the OSM Semantic Network and WordNet synsets, we obtain a human-generated mapping M_h, which includes 114 correct mappings. This dataset can be utilised as a ground truth to evaluate *Voc2WordNet*, our semantic mapping technique.

5.2 Evaluation Measures

To evaluate the performance of *Voc2WordNet*, we define the following performance measures. Following Euzenat [14], we assume that a correct mapping m belongs both to the machine mapping M and the human mapping M_h ($m \in M \wedge m \in M_h$). By contrast, an incorrect mapping only belongs to the machine mapping ($m \in M \wedge\ m \notin M_h$). Hence, we define precision P and recall R of mapping M as:

$$P_M = \frac{|M \cap M_h|}{|M|} \quad R_M = \frac{|M \cap M_h|}{|M_h|} \quad P_M, R_M \in [0,1] \tag{2}$$

All these measures fall in the interval $[0,1]$, with 1 as the best possible result ($M \equiv M_h$), and 0 as the worst ($M \cap M_h = \emptyset$). These measures will be used as indicators of the quality of the semantic mapping in the next sections.

5.3 Experiment Set-Up

The algorithm *Voc2WordNet* takes five parameters: V, W, ol_{min}, f_{min}, and Θ (see Section 4). Keeping the vocabulary V and WordNet W constant, we want to assess the impact of the other three parameters, ol_{min}, f_{min}, and Θ. Hence, we define the following parameters:

- Salient taxonomy Θ: either $\Theta \equiv W$ (i.e. taxonomy disabled), or a taxonomy of geographic terms (2 options);
- Minimum lexical overlap ol_{min}: $\{0,1,2\}$ (3 options);
- Minimum word sense frequency f_{min}: $\{0,1,2\}$ (3 options).

These parameters result in 18 unique combinations of parameters. A random disambiguation approach is added as a baseline. In order to disambiguate the terms from the OSM Semantic Network to the corresponding word sense in WordNet synsets, we select a subset of the WordNet taxonomy Θ that is relevant to the OSM context, i.e. entities and processes that are employed to describe OSM objects.

By manually observing the upper level of WordNet (i.e. synsets with depth ≤ 3), we selected eight synsets as roots of the salient taxonomy (see Table 2). All children synsets were subsequently recursively extracted, resulting in a salient taxonomy Θ of 6,312 noun synsets, navigating the *wns:hyponymOf* and *wns:partMeronymOf* relations. The salient taxonomy corresponds to about 7% of the entire WordNet noun taxonomy. The algorithm *Voc2WordNet* was executed on the 18 parameter combinations.

Table 2. Salient synsets in the upper part of the WordNet taxonomy

Salient taxonomical roots in WordNet	
wn:location-noun-1	*wn:artifact-noun-1*
wn:land-noun-2	*wn:activity-noun-1*
wn:ecosystem-noun-1	*wn:water_system-noun-1*
wn:natural_object-noun-1	*wn:natural_phenomenon-noun-1*

5.4 Experiment Results

The experiment generated 18 mappings of the OSM Semantic Network on WordNet synsets. Each mapping was compared with the human-generated dataset described in Section 5.1, obtaining precision and recall values. In order to analyse the impact of each parameter on the results, we summarise the performance indicators in Table 3, showing the mean precision $\bar{P}_M$ and recall $\bar{R}_M$. As expected, precision and recall are inversely proportional. All of the three filters (Θ, f_{min}, ol_{min}) have a positive impact on the precision, and a negative impact on the recall. The filter based on the salient taxonomy Θ improves the mean precision $\bar{P}_M$ from .72 to .81, with a minimal loss of recall. Similarly, the filter based on f_{min} and ol_{min} increases the mean precision at the expense of the mean recall. These results support the validity of the key ideas behind *Voc2WordNet*, described in Section 4.

Table 3. Experiment results of *Voc2WordNet* on the OSM Semantic Network. (*) Best precision and recall.

Parameter name	Parameter value	Mean $\bar{P}_M$	Mean $\bar{R}_M$
Random baseline	–	.21	.34
Taxonomy Θ	*off*	.79	.5*
	on	.88*	.49
Min frequency f_{min}	*(off)* 0	.82	.56
	1	.84*	.56
	2	.84*	.54
Min lexical overlap ol_{min}	*(off)* 0	.7	.82*
	1	.75	.81
	2	.87*	.49
Upper bounds	–	.88	.82

Considering the upper bounds obtained in this preliminary experiment ($P = .88$, $R = .82$), we consider *Voc2WordNet* to be a promising approach to grounding a vocabulary such as the OSM Semantic Network in WordNet. The optimal choice of the three parameters largely depends on the specific context in which *Voc2WordNet* is being applied. Based on specific users' needs, precision could be favoured over recall, or vice-versa. In order to extend this initial evaluation further, more terms could be included in the dataset, and the manual mapping could be performed and validated by a group of independent human subjects. In addition, the optimal parameters could be obtained using machine learning techniques on a desired training set of mappings.

6 Conclusions

Linked open data (LOD) constitutes a promising paradigm to create a shared semantic space, in which heterogenous geospatial datasets can inter-operate. In

the LOD cloud, WordNet can be used as a shared semantic ground to enable inter-operability between heterogenous vocabularies. In this paper, we described our two-fold contribution to the LOD cloud. First, we described the structuring of the OSM Semantic Network as LOD, using the W3C Simple Knowledge Organization System (SKOS). Second, we outlined and evaluated a semantic mapping algorithm, *Voc2WordNet*, which aimed at mapping a given vocabulary onto WordNet. The following conclusions can be drawn:

- The OSM Semantic Network bridges the semantics of OSM data and the LOD cloud. The network is extracted from the OSM Wiki website, a repository where contributors define, edit, and document the semi-structured folksonomy of tags. The dataset is structured as a SKOS vocabulary of terms utilised to describe OSM geographic features. We made the OSM Semantic Network freely available online,[8] and we linked it to existing semantic resources, including LinkedGeoData and TagInfo.
- Despite the advances reported in this article, the OSM Semantic Network presents a number of open challenges. As happens with crowdsourced resources, the network inevitably contains some degree of noise, ambiguity, and incorrect semantic mappings. Being a folksonomy, the OSM Semantic Network does not necessarily reflect ontological commitments in the vector data, and should therefore be utilised taking into account the intrinsic uncertainty of VGI.
- Our algorithm *Voc2WordNet* offers a general semantic mapping technique between a specialised vocabulary and the well-known lexical database WordNet. Given an input term from the vocabulary, *Voc2WordNet* identifies salient synsets in WordNet using three salience indicators: (1) the usage frequency of a term; (2) the term overlap between the lexical definition of the given term and the WordNet definition; and (3) a manually selected salient taxonomy. These indicators can be combined to increase precision, with a minor loss in recall. *Voc2WordNet* was tested on the OSM Semantic Network, obtaining high precision (.88) and recall (.82). A more extensive evaluation is necessary to demonstrate the effectiveness of *Voc2WordNet* across different vocabularies.

The OSM Semantic Network provides general-purpose semantic support for exploiting OSM data in geo-applications. Its integration with LinkedGeoData and WordNet enables the discovery of implicit semantic relations between map features, e.g. subsumption or meronomy, as well as the discovery of affordances, a promising approach to modelling the role of places. The network can support a number of semantic tasks, facilitating the computation of semantic similarity of geographic terms, and the matching of the same entities across LinkedGeoData, DBpedia, GeoNames, and other geo-knowledge bases [6].

Similarly, using GeoSPARQL [8] and federated queries over the LOD cloud,[9] it is possible, for example, to retrieve the schools from LinkedGeoData within

[8] `http://wiki.openstreetmap.org/wiki/OSMSemanticNetwork` (acc. Oct 30, 2012).
[9] `http://www.w3.org/TR/sparql11-federated-query` (acc. Oct 30, 2012).

a given geographic location, and to use the OSM Semantic Network to perform a semantic query expansion to features semantically related to school, such as kindergardens, highschools, and colleges.

Structuring VGI according to the LOD paradigm provides a valuable contribution to deliver richer, more structured geospatial information to both humans and machines. However, the LOD cloud presents a number of limitations that need to be addressed, in particular in relation to the management of identity [20], and spatio-temporal reasoning [21]. These issues notwithstanding, the LOD cloud already provides an open laboratory to a growing community of scientists, software developers, and GIS specialists. The OSM Semantic Network and *Voc2WordNet* constitute two further steps towards the inclusion of VGI into this vast semantic space.

References

1. Ashish, N., Sheth, A. (eds.): Geospatial Semantics and the Semantic Web: Foundations, Algorithms, and Applications, vol. 12. Springer, New York (2011)
2. Auer, S., Lehmann, J., Hellmann, S.: LinkedGeoData: Adding a Spatial Dimension to the Web of Data. In: Bernstein, A., Karger, D.R., Heath, T., Feigenbaum, L., Maynard, D., Motta, E., Thirunarayan, K. (eds.) ISWC 2009. LNCS, vol. 5823, pp. 731–746. Springer, Heidelberg (2009)
3. Baglatzi, A., Kokla, M., Kavouras, M.: Semantifying OpenStreetMap. In: Proceedings of the 5th International Terra Cognita Workshop 2012 - Foundations, Technologies and Applications of the Geospatial Web. CEUR Workshop Proceedings, vol. 901, pp. 39–50 (2012)
4. Ballatore, A., Bertolotto, M.: Semantically Enriching VGI in Support of Implicit Feedback Analysis. In: Kim, K.-S. (ed.) W2GIS 2011. LNCS, vol. 6574, pp. 78–93. Springer, Heidelberg (2010)
5. Ballatore, A., Bertolotto, M., Wilson, D.: Geographic Knowledge Extraction and Semantic Similarity in OpenStreetMap. In: Knowledge and Information Systems, pp. 1–21 (2012)
6. Ballatore, A., Wilson, D., Bertolotto, M.: A Survey of Volunteered Open Geo-Knowledge Bases in the Semantic Web. In: Advanced Techniques in Web Intelligence - 3: Quality-based Information Retrieval. SCI. Springer (2012) (in Press)
7. Ballatore, A., Wilson, D.C., Bertolotto, M.: The Similarity Jury: Combining Expert Judgements on Geographic Concepts. In: Castano, S., Vassiliadis, P., Lakshmanan, L.V.S., Lee, M.L. (eds.) ER Workshops 2012. LNCS, vol. 7518, pp. 231–240. Springer, Heidelberg (2012)
8. Battle, R., Kolas, D.: Enabling the Geospatial Semantic Web with Parliament and GeoSPARQL. Semantic Web 3(4), 355–370 (2012)
9. Berners-Lee, T.: Linked Data (2006), `http://www.w3.org/DesignIssues/LinkedData.html`
10. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. Scientific American 284(5), 28–37 (2001)
11. Bizer, C., Heath, T., Berners-Lee, T.: Linked Data – The Story So Far. International Journal on Semantic Web and Information Systems 5(3), 1–22 (2009)

12. Coast, S.: OpenStreetMap - The Best Map (February 19, 2010), OpenGeoData http://opengeodata.org/openstreetmap-the-best-map
13. Elwood, S., Goodchild, M., Sui, D.: Researching Volunteered Geographic Information: Spatial Data, Geographic Research, and New Social Practice. Annals of the Association of American Geographers 102(3), 571–590 (2012)
14. Euzenat, J.: Semantic precision and recall for ontology alignment evaluation. In: Proc. 20th International Joint Conference on Artificial Intelligence (IJCAI), pp. 348–353 (2007)
15. Euzenat, J., Meilicke, C., Stuckenschmidt, H., Shvaiko, P., Trojahn, C.: Ontology Alignment Evaluation Initiative: Six Years of Experience. In: Spaccapietra, S. (ed.) Journal on Data Semantics XV. LNCS, vol. 6720, pp. 158–192. Springer, Heidelberg (2011)
16. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database. MIT Press, Cambridge (1998)
17. Gangemi, A., Guarino, N., Masolo, C., Oltramari, A., Schneider, L.: Sweetening Ontologies with DOLCE. In: Gómez-Pérez, A., Benjamins, V.R. (eds.) EKAW 2002. LNCS (LNAI), vol. 2473, pp. 166–181. Springer, Heidelberg (2002)
18. Giunchiglia, F., Maltese, V., Farazi, F., Dutta, B.: GeoWordNet: A Resource for Geo-spatial Applications. In: Aroyo, L., Antoniou, G., Hyvönen, E., ten Teije, A., Stuckenschmidt, H., Cabral, L., Tudorache, T. (eds.) ESWC 2010, Part I. LNCS, vol. 6088, pp. 121–136. Springer, Heidelberg (2010)
19. Goodwin, J., Dolbear, C., Hart, G.: Geographical Linked Data: The Administrative Geography of Great Britain on the Semantic Web. Transactions in GIS 12, 19–30 (2008)
20. Halpin, H., Hayes, P.J., McCusker, J.P., McGuinness, D.L., Thompson, H.S.: When owl:sameAs Isn't the Same: An Analysis of Identity in Linked Data. In: Patel-Schneider, P.F., Pan, Y., Hitzler, P., Mika, P., Zhang, L., Pan, J.Z., Horrocks, I., Glimm, B. (eds.) ISWC 2010, Part I. LNCS, vol. 6496, pp. 305–320. Springer, Heidelberg (2010)
21. Janowicz, K., Scheider, S., Pehle, T., Hart, G.: Geospatial Semantics and Linked Spatiotemporal Data: Past, Present, and Future. In: Semantic Web – Special Issue on Linked Spatiotemporal Data and Geo-Ontologies, pp. 1–13 (2012)
22. Lin, H., Davis, J., Zhou, Y.: An Integrated Approach to Extracting Ontological Structures from Folksonomies. In: Aroyo, L., Traverso, P., Ciravegna, F., Cimiano, P., Heath, T., Hyvönen, E., Mizoguchi, R., Oren, E., Sabou, M., Simperl, E. (eds.) ESWC 2009. LNCS, vol. 5554, pp. 654–668. Springer, Heidelberg (2009)
23. Mendes, P., Jakob, M., García-Silva, A., Bizer, C.: DBpedia Spotlight: Shedding Light on the Web of Documents. In: Proceedings of the 7th International Conference on Semantic Systems, pp. 1–8. ACM (2011)
24. Miles, A., Matthews, B., Wilson, M., Brickley, D.: SKOS Core: Simple Knowledge Organisation for the Web. In: International Conference on Dublin Core and Metadata Applications, DC-2005, pp. 3–10. DCMI Publications (2005)
25. Mooney, P., Corcoran, P.: Characteristics of heavily edited objects in OpenStreetMap. Future Internet 4(1), 285–305 (2012)
26. Navigli, R.: Word sense disambiguation: A survey. ACM Computing Surveys 41(2), 1–10 (2009)
27. Noy, N.: Semantic Integration: A Survey Of Ontology-Based Approaches. SIGMOD Record 33(4), 65–70 (2004)
28. Purves, R., Jones, C.: Geographic Information Retrieval. SIGSPATIAL Special 3(2), 2–4 (2011)

29. Ramage, D., Rafferty, A., Manning, C.: Random walks for text semantic similarity. In: Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing, pp. 23–31. ACL (2009)
30. Singhal, A.: Introducing the Knowledge Graph: things, not strings (May 16, 2012), http://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html
31. Suominen, O., Hyvönen, E.: Improving the Quality of SKOS Vocabularies with Skosify. In: ten Teije, A., Völker, J., Handschuh, S., Stuckenschmidt, H., d'Acquin, M., Nikolov, A., Aussenac-Gilles, N., Hernandez, N. (eds.) EKAW 2012. LNCS, vol. 7603, pp. 383–397. Springer, Heidelberg (2012)
32. Vander Wal, T.: Folksonomy (2007), http://vanderwal.net/folksonomy.html