Linking Geographic Vocabularies through WordNet

A. Ballatore,[*] M. Bertolotto,[†] and D.C. Wilson[‡]

Author copy. Published in *Annals of GIS*, 20 (2) 2014

Abstract

The linked open data paradigm has emerged as a promising approach to structuring and sharing geospatial information. One of the major obstacles to this vision lies in the difficulties found in the automatic integration between heterogeneous vocabularies and ontologies that provides the semantic backbone of the growing constellation of open geo-knowledge bases. In this article, we show how to utilise WordNet as a semantic hub to increase the integration of linked open data. With this purpose in mind, we devise *Voc2WordNet*, an unsupervised mapping technique between a given vocabulary and WordNet, combining intensional and extensional aspects of the geographic terms. *Voc2WordNet* is evaluated against a sample of human-generated alignments with the OpenStreetMap Semantic Network, a crowdsourced geospatial resource, and the GeoNames ontology, the vocabulary of a large digital gazetteer. These empirical results indicate that the approach can obtain high precision and recall.

Keywords: Geo-semantics, Linked open data, OSM Semantic Network, SKOS, GeoNames, WordNet, OpenStreetMap, Semantic integration, Semantic mapping, LIMES, Voc2WordNet

1 Introduction

Over the past decades, a large volume of digital information has been disseminated online in a variety of incompatible formats and heterogeneous data spaces. This semantic gap hinders the ability to analyse, explore, and discover unexpected connections and relations between entities, obtaining insights about complex social, geographic, cultural, and economic processes. Berners-Lee's Semantic Web is a prominent attempt to overcome this crucial gap, and to provide

[*] School of Computer Science and Informatics, University College Dublin, Ireland. andrea.ballatore@ucd.ie

[†] School of Computer Science and Informatics, University College Dublin, Ireland.

[‡] Department of Software and Information Systems, University of North Carolina, Charlotte, NC

a flexible and yet unified platform for data sharing (Berners-Lee et al., 2001). One of the most promising initiatives in this ambitious framework 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 (i) released under open licenses; (ii) saved in a machine-readable digital format; (iii) stored in non-proprietary formats; (iv) accessible via URIs; and (v) linked to other LOD.[1] As LOD is generated and published online, a graph of datasets has emerged, resulting in the LOD cloud, also referred to as the Web of Data, in which hundreds of diverse data sources enjoy varying degrees of semantic integration through links, with a variety of access points (Bizer et al., 2009).[2]

As a large part of online data involves a spatial dimension, geographic entities and their semantics play a central role in the LOD cloud, facilitating the geospatial grounding of scientific and commercial data (Hart and Dolbear, 2013; Janowicz et al., 2012). The LOD paradigm is promising in the context of geographic information retrieval, where existing techniques have shown limited effectiveness (Purves and Jones, 2011). For example, the LOD-based search engine Wikipedia Faceted Search handled complex geospatial queries, e.g. ‘Which Rivers flow into the Rhine and are longer than 50 kilometers?’ (Hahn et al., 2010). The emergence of the LOD infrastructure also has 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 informational assets as LOD[3] (Goodwin et al., 2008).

To enable the promising network effects in the LOD cloud, datasets need to be inter-connected through meaningful relationships. Generating such semantic mappings automatically is therefore a crucial part of the LOD vision, enabling interoperability while preserving local semantic details. In the LOD jargon, the process of linking a new dataset to existing ones is called ‘bootstrapping,’ and is usually performed on semantic hubs such as DBpedia (Mendes et al., 2011). In this article, extending a preliminary study (Ballatore et al., 2013b), we focus on the bootstrapping of geographic vocabularies, utilising WordNet as a LOD hub.

In this context, we first describe *Voc2WordNet*, a generic technique to generate a semantic mapping between a given vocabulary and WordNet, which we selected as a shared semantic ground because of its rich relations (Fellbaum, 2010). This semantic mapping is valuable because it can support and enable a number of natural language processing and information retrieval operations on geographic LOD. *Voc2WordNet* is aimed at the underspecified vocabularies adopted in geo-knowledge bases, to increase their interoperability, and to enable the discovery of rich ontological relations such as part-whole (e.g. part-of relations) and subsumption (e.g. is-a relations), which are present in WordNet. Second, we evaluate *Voc2WordNet* on two real datasets containing primarily geographic information, the crowdsourced OSM Semantic Network and the lightweight GeoNames ontology which provides a vocabulary to a large dig-

[1] `http://5stardata.info` - All URLs cited were accessed on April 21, 2014.

[2] See for example `http://thedatahub.org`

[3] `http://data.ordnancesurvey.co.uk`

ital gazetteer.

The remainder of this article is organised as follows. Section 2 reviews relevant work in the areas of LOD integration, open geo-knowledge bases, geo-semantics, and WordNet. This section also describes the OSM Semantic Network and the GeoNames ontology, which are used in the evaluation. Section 3 describes and formalises *Voc2WordNet*, a generic approach to semantic mapping onto WordNet. Subsequently, we report on the evaluation of the approach, executed on a sample of terms from the OSM Semantic Network and the GeoNames ontology, and compared with existing LOD mapping tools in Section 4. Finally, conclusions and directions for future research are discussed in Section 5.

2 Related work

The approach to LOD integration proposed in this article is inscribed in the Semantic Geospatial Web research, in which identification of the same concepts and entities in heterogeneous data spaces through semantic similarity measures is considered to be a crucial enabler (Janowicz et al., 2012). More generally, the automatic merging of different conceptual schemas is a time-honoured challenge in computer science, beginning well before the advent of the Semantic Web. Two datasets can be aligned at the schema level (e.g. matching the concept 'river' in both ontologies), and at the instance level (e.g. connecting the Po River in both knowledge bases). Logical reasoning, machine learning, and statistical analysis have been utilised to tackle the problem in the context of database schemas (Noy, 2004). Since 2005, the Ontology Alignment Evaluation Initiative (OAEI) has proposed benchmarks and performance metrics specifically tailored to the area of ontology alignment and integration (Euzenat et al., 2011).

Several approaches to generate 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, semantic methods can observe the terms from multiple angles: *internal* methods observe aspects of the terms in isolation, such as 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. As covered in the next section, these approaches are utilised in actual information integration software tools.

2.1 LOD integration frameworks

To perform the integration of LOD datasets stored in RDF format, a number of frameworks have been developed. The RDF-AI tool aims at the integration of RDF datasets (Scharffe et al., 2009). The matching is performed by computing the semantic similarity of two given entities, based on a user-provided set of salient properties (e.g. the title and year of a musical work, the author and title of a book, etc.). The semantic similarity can be computed either with fuzzy string matching based on the sequence integration algorithm, or by comparing synonyms in WordNet. Subsequently RDF-AI uses the matching pairs either to fuse two datasets into one, or to generate a list of matching entities.

Along similar lines, Volz et al. (2009) developed *Silk Link Discovery Framework*, which aims at establishing relations between entities in different data sources. A number of strategies can be used to match properties, based on simple string similarity measures. The user can specify what properties should be compared and with which similarity metric, and can specify the thresholds above which the relations should be established or should be manually verified. For example, in a given context, all pairs with similarity equal to or greater than 0.9 might be linked automatically, while pairs with similarity greater than 0.6 but smaller than 0.9 should be checked manually. Such heuristics can be defined in the Link Specification Language (Silk-LSL). More recently, Isele and Bizer (2012) extended Silk with the *GenLink* algorithm, which extracts rules from valid links using supervised machine learning.

Scalability issues affect these tools, which often are crippled by the enormous complexity of the brute-force comparison of large datasets. To overcome this issue, Ngomo and Auer (2011) developed the *LInk discovery framework for MEtric Spaces* (LIMES). This framework performs operations logically equivalent to those of Silk, but relies on the concept of triangle inequality in metric spaces to compute pessimistic estimates of instance similarities. Based on these approximations, LIMES can exclude a large number of entity pairs that cannot satisfy the user-defined matching conditions. The actual similarities of the remaining pairs are then computed and the matching instances are returned, without losing recall. While these frameworks are useful in the context of a generic matching between entities in LOD datasets, they do not perform well in the case of WordNet, as discussed in Section 4.3.

2.2 WordNet as a semantic hub

Since the early 1990s, WordNet has been a valuable semantic resource for many applications in natural language processing and artificial intelligence (Fellbaum, 1998, 2010). The core element of WordNet is the 'synset,' a concept that aggregates a set of synonymous words, called 'word senses.' For example, the geographic concept 'stream' is represented in WordNet by synset {*stream,watercourse*}. This synset contains two word senses, *stream#n#1* and *watercourse#n#1*, with the notation *word#part-of-speech#word-sense-number*. The word 'stream' appears in five different synsets, capturing its high polysemy. Synsets are connected through several semantic relations, such as *similarTo*, *partMeronymOf*, *adjectivePertainsTo*, *causes*, *antonymOf*, and *entails.*[4] Two versions of WordNet, 2.0 and 3.0, are currently linked in the LOD cloud.[5]

[4] See `http://www.w3.org/2006/03/wn/wn20/schemas/wnfull.rdfs` for the complete list.

[5] `http://www.w3.org/2006/03/wn/wn20` and `http://semanticweb.cs.vu.nl/lod/wn30`

WordNet has found particular success in the areas of word sense disambiguation and semantic similarity (Navigli, 2009; Ballatore et al., 2012). 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. Ramage et al., 2009). Although the semantic network was not designed for this purpose, it has been frequently used as a general-purpose semantic ground, for example to discover semantic connections in unstructured data (Lin et al., 2009). The limitations of WordNet have been thoroughly discussed. Being a top-down, expert-controlled resource, its lexical coverage is bound to be lower than that of crowdsourced alternatives, such as DBpedia. Furthermore, the upper part of its taxonomical structure has been critised as ontologically unsound, prompting a substantial re-design and refinement, following state-of-the-art ontological theories (Gangemi et al., 2003).

A large number of projects provide WordNet-like semantic networks in languages other than English.[6] To date, none of the numerous alternative semantic resources has yet managed to dethrone WordNet from its leading position as general-purpose semantic ground. In the context of the LOD cloud, WordNet has been used as a high-quality primary semantic source in many projects interlinked with DBpedia, the largest hub of the LOD cloud (Ballatore et al., 2013). Although DBpedia has considerably larger coverage than WordNet, its ontological structure is lighter, and provides fewer semantic relations. For this reason, we argue that WordNet could complement DBpedia as a central resource in the LOD cloud. Using WordNet as an imperfect, and yet rich semantic ground, it is possible to integrate geo-vocabularies, such as the OSM Semantic Network and the GeoNames ontology, described in the next sections.

2.3 The OSM Semantic Network

Volunteered geographic information (VGI) is playing an increasingly important role in the LOD cloud. From its foundation in 2004, OpenStreetMap (OSM) has established itself as the most ambitious VGI project. The OSM conceptualisation emerges from semantic negotiations within the contributors' community, reaching consensus around the intended meaning and usage of 'tags,' i.e. terms describing geographic entities. This radically open approach to geosemantics was adopted by the project's creators on the assumption 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. The downside of the adoption of a semi-structured folksonomy is, predictably, wide variability and ambiguity in the terms' interpretation, proliferation of near-synonym terms, and lack of explicit semantic relations (Ballatore and Bertolotto, 2011). The OSM Semantic Network is interlinked with LinkedGeoData and DBpedia (Auer et al., 2009). Using *Voc2WordNet*, described in Section 3, the network has also been linked to WordNet.

To provide a knowledge-based support tool for OSM, we extracted the OSM

[6] See the list at `http://www.globalwordnet.org/gwa/wordnet_table.html`

Semantic Network, a semantic artefact containing the conceptualisation of OSM tags, providing a machine-readable structure that can support the automatic manipulation of OSM features in data mining, geographic information retrieval, and information integration (Ballatore et al., 2013b).[7] The network was initially developed offline to compute the semantic similarity of tags (Ballatore et al., 2013a), and is published in the LOD cloud.[8] The OSM Semantic Network is organised as a W3C Simple Knowledge Organization System (SKOS) vocabulary (Miles et al., 2005). 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 in SKOS 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*). A concept is described by a preferred short lexical label (*skos:prefLabel*), and can have n alternative labels (*skos:altLabel*). A more extensive and unique definition can be given to a concept in a given language (*skos:definition*). Hence, each term defined in the network corresponds to a SKOS concept. For example, the OSM tag *waterway=river* corresponds to the term *osnt:k:waterway/v:river*.[9] The quality of the SKOS vocabulary was assessed based on the criteria outlined by Suominen and Hyvönen (2012). Another example of a SKOS-based vocabulary is the GeoNames ontology, described in the next section.

2.4 The GeoNames ontology

The GeoNames project is an open digital gazetteer combining a variety of data sources, representing the location of about 8 million unique features.[10] Thanks to its impressive coverage, this gazetteer is widely used in geospatial applications, and constitutes a densely linked resource in the LOD cloud. The geographic features contained in GeoNames are classified using a simple hierarchical tree, in which 9 Feature Classes (e.g. *Populated places*) contain more specific 690 Feature Codes (e.g. *religious populated places*). Although this artefact is a lightweight SKOS vocabulary with little formal ontological content, it is referred to as the GeoNames ontology, and has reached version 3.1.

The peculiarities and issues found in the GeoNames ontology have been discussed by Giunchiglia et al. (2010), who integrated it manually with WordNet to generate GeoWordNet, a geographically enhanced version of WordNet. Although this integration provides indeed a useful resource, our contention is that automated interlinking should be preferred to the manual semantic merging applied in GeoWordNet. Even if automated semantic bootstrapping is unlikely to equal manual mapping in terms of quality, it provides a sustainable way to

[7] `http://wiki.openstreetmap.org/wiki/OSMSemanticNetwork`

[8] `http://datahub.io/dataset/osm-semantic-network`

[9] `http://spatial.ucd.ie/lod/osn/term/k:waterway/v:river`

[10] `http://www.geonames.org`

Symbol	Description
V	Vocabulary, i.e. set of terms t. E.g. the GeoNames ontology
t	Generic term $\in V$. E.g. *osnt:k:waterway*
Θ	Salient taxonomy extracted from WordNet.
W	WordNet, i.e. a set of synsets.
s	WordNet synset, $s \in W$. E.g. *wn:river-noun-1*
ws	Word sense in synset s. E.g. *wn:wordsense-river-noun-1*
C_t	Candidate synsets $s \in W$ for term t
$ol(t,s)$	Overlap between definitions of term t and synset s. $ol \geq 0$
$f(ws)$	Usage frequency of $ws \in s$. $f \geq 0$
ol_{min}	Minimum lexical overlap between terms.
f_{min}	Minimum frequency of word sense in WordNet.
$\sigma(s, ws, t)$	Salience score for candidate s and ws for term t.
$M(V,W)$	Set of semantic mappings m between vocabulary V and W
m	Semantic mapping $< t, r, s >$ between term $t \in V$ and synset $s \in W$, with relation r
r	Relation that defines the nature of the semantic mapping m: exact, close, or related (see Section 3.1)

Table 1: Notations

include new resources in the LOD cloud, without increasing the fragmentation of existing resources into multiple versions and preserving the structure of each resource and their local semantics.

In this sense, whilst GeoWordNet is the result of a merging process, resulting in a new resource, *Voc2WordNet* provides an automatic mapping technique between a given vocabulary and WordNet. 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, and *Voc2WordNet* has precisely the purpose of filling this specific gap. In this sense, it is not a general-purpose ontology mapping technique. As described in the next section, *Voc2WordNet* performs the semantic mapping between a vocabulary term and a specific WordNet word sense both from an intensional (i.e. lexical overlap between the lexical definitions) and an extensional perspective (i.e. the usage frequency).

3 *Voc2WordNet*, a semantic mapping algorithm

To increase integration and interoperability of linked open data (LOD) at the schema level, we propose to utilise the lexical database WordNet as a semantic hub. For this purpose, this section describes *Voc2WordNet*, an algorithm devised to generate a semantic mapping between a given vocabulary and 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. The issue tackled by *Voc2WordNet* is inscribed within the open problem of word sense disambiguation, i.e. distinguishing when

Abbr.	Description	URI
rdfs	RDF schema	`http://www.w3.org/2000/01/rdf-schema#`
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/`
osn	OSM Semantic Network	`http://spatial.ucd.ie/lod/osn/`
osnt	− tag	`http://spatial.ucd.ie/lod/osn/term/k:<key>/v:<value>`
osnpt	− proposed term	`http://spatial.ucd.ie/lod/osn/proposed_term/`
gno	GeoNames ontology	`http://www.geonames.org/ontology#`
lgdo	LinkedGeoData	`http://linkedgeodata.org/ontology/`

Table 2: XML namespaces

the word 'bank' refers to a financial institution or to the terrain alongside a river (Navigli, 2009). The similarity notwithstanding, the constraints in which *Voc2WordNet* operates make the integration considerably simpler than open word sense disambiguation on raw text.

The *Voc2WordNet* approach is primarily aimed at the schema level typical of vocabularies, and not at the instance level, and combines intensional and extensional aspects to identify salient synsets in WordNet. Although this article focuses on geo-vocabularies, *Voc2WordNet* can be used to map any vocabulary into WordNet. The notations used in the remainder of this article are reported in Table 1. For the sake of brevity, the namespaces are summarised in Table 2. Section 3.1 defines the nature and scope of the semantic mapping for which *Voc2WordNet* is designed. The detailed workings of *Voc2WordNet* are subsequently described in Section 3.2.

3.1 Mapping relations

A semantic mapping m between term $t \in V$ and synset $s \in W$ has the form $< t, r, s >$. Given the aim of SKOS to provide a Web and collaborative platform for vocabularies, the language provides semantic relations to connect concepts to equivalent, similar or related concepts in other vocabularies. Such relations are called *mapping properties*.[11] A concept can engage in an identity relation with a concept in another schema (*skos:exactMatch*), can be very similar (*skos:closeMatch*), or can be only loosely related to it (*skos:relatedMatch*). In the context of *Voc2WordNet*, we adopt three SKOS symmetric mapping relations r:

Related (*skos:relatedMatch*): General semantic relatedness (e.g. *osnt:k:power/-v:station* and *wn:electricity-noun-1*);

Close (*skos:closeMatch*): Highly similar terms which originated from different information communities (e.g. *osnt:k:wood* and *wn:forest-noun-2*);

Exact (*skos:exactMatch*): Terms that originated from the same information community, but expressed in different vocabularies (e.g. *osnt:k:amenity/-*

[11] `http://www.w3.org/TR/skos-reference/#mapping`

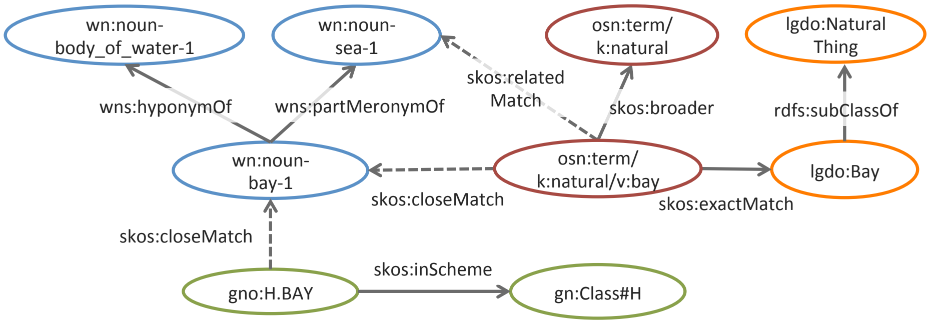

Figure 1: Fragments of entities representing geographic concept 'bay' and their mappings in WordNet (*wn*), LinkedGeoData (*lgdo*), the OSM Semantic Network (*osn*), and the GeoNames ontology (*gno*). Dotted relations are generated by *Voc2WordNet*.

v:university and *lgdo:University*). We consider this mapping to be logically equivalent to *owl:sameAs*.

Through these relations, it is possible to establish a mapping $m =< t, r, s >$ between the vocabulary V and the WordNet synsets W. We define the validity of a mapping in terms of its semantic coherence (is the mapping's semantics clear to a human observer?) and completeness (does the mapping include all the possible coherent relationships?). Figure 1 shows a fragment of a possible valid mapping of the geographic term 'bay' between the GeoNames ontology, the OSM Semantic Network, LinkedGeoData, and WordNet. To further illustrate the difficulties of the semantic mapping with WordNet, the definition of *wn:bay-noun-1* is "an indentation of a shoreline larger than a cove but smaller than a gulf," while *wn:bay-noun-2* is defined as "the sound of a hound on the scent," an alternative and semantically unrelated meaning. The OSM term *osnt:k:natural/v:bay* is defined as a "a large body of water partially enclosed by land but with a wide mouth." The following list shows possible correct and incorrect mappings between these terms:

(a) <*osnt:k:natural/v:bay close wn:bay-noun-1*> (correct)

(b) <*osnt:k:natural/v:bay related wn:sea-noun-1*> (correct)

(c) <*osnt:k:natural/v:bay related wn:bay-noun-1*> (incorrect)

(d) <*osnt:k:natural/v:bay related wn:bay-noun-2*> (incorrect)

(e) <*osnt:k:natural/v:bay close wn:sea-noun-1*> (incorrect)

Case (e) should considered incorrect because the synset 'sea' is only *related* to 'bay,' and does not constitute a close match. In some situations, the distinction between *close* and *related*, and *close* and *exact*, is more nuanced, and both cases can be considered correct.

3.2 Algorithm

Voc2WordNet generates a mapping 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 the lexical definition of t, taking compound words into account (e.g. 'swimming pool'), and then splitting them if not defined directly 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 the lexical definitions of 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 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 $\sigma = .8$.

These three indicators are combined to select valid mappings both from the term itself t, and from the term's lexical definition, which can contain useful pointers to relevant terms (e.g. the definition of term 'power station' contains 'electricity'). In order to provide more leverage, the algorithm filters out candidates based on a minimum frequency (f_{min}), a minimum overlap (ol_{min}), and a manually selected salient taxonomy (Θ). The detailed workings of the algorithm and functions are outlined in Algorithm 1. In the next section, *Voc2WordNet* is evaluated on two real-world datasets, i.e. the OSM Semantic Network and the GeoNames ontology.

4 Evaluation

This section describes an experimental evaluation of *Voc2WordNet*, our semantic mapping technique, outlined in Section 3, which extends an initial exploration of the algorithm (Ballatore et al., 2013b). We generated two evaluation datasets M_h by selecting random samples of terms from the OSM Semantic Network and the GeoNames ontology (Section 4.1). To measure the performance of the algorithm, we defined performance measures (precision, recall, and an F-measure) that compare the machine-generated mapping M with the human mapping M_h (Section 4.2). In order to compare *Voc2WordNet* with existing tools, preliminary experiments were conducted on the mapping framework LIMES (Section 4.3). Finally, an experiment on a number of parameter combinations was executed on both datasets (Section 4.4), and the performance of *Voc2WordNet* is analysed and discussed (Section 4.5).

4.1 Evaluation datasets

To construct a gold standard for this evaluation, we selected a random sample of 30 terms from the OSM Semantic Network (see Section 2.3) and 30 terms from the GeoNames ontology (see Section 2.4). This random sample corresponds to approximately 1% of terms in OSM Semantic Network, and to 4% of terms in the GeoNames ontology. The sample terms were manually mapped to semantically

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 $\Theta$
output: Set $M$ of semantic mappings $m =< t, r, s >$

$M \leftarrow \emptyset$
foreach term $t \in V$ do
    $m \leftarrow$ findSemanticMapping$(t, W)$;
    add $m$ to $M$;
    extract terms from lexical definition of $t$ to set $D_t$;
    foreach term $d \in D_t$ do
        $m_d \leftarrow$ findSemanticMapping$(d, W)$
        set ‘related’ as $r$;
        add $m_d$ to $M$;
return $M$.
```

Function findSemanticMapping(t, W)

```
$C_t \leftarrow \emptyset$
foreach $ws \in W$ do
    find set of matching word senses $ws \in W$ with lexicalMatch$(ws, t)$;
    find synset $s$ corresponding to $ws$ in WordNet;
    if $s \notin \Theta$, skip $ws$;
    fetch word sense frequency $f(ws)$ from WordNet;
    if $f(s) < f_{min}$, skip $ws$;
    compute lexical overlap between definitions $ol(s, t)$;
    if $ol(s, t) < ol_{min}$, skip $ws$;
    $s$ and $ws$ are a valid candidate, add pair $< s, ws >$ to candidate set
    $C_t$;
foreach $< s, ws > \in C_t$ do
    compute salience score $\sigma(s, ws, t)$;
select best candidate $s_b \in C_t$ having $max(\sigma(s, ws, t))$;
if lexicalMatch$(ws, t)$ is ‘complete’ $\wedge$ $max(ol(s, t)) \wedge max(f(ws))$ then
    select ‘close’ as $r$
else
    select ‘related’ as $r$
generate mapping $m =< t, r, s_b >$ and return it.
```

```
Function lexicalMatch($ws, t$)
if $ws$ is contained in $t$ then
    return ‘partial’;
if $ws$ is equal to $t$ then
    return ‘complete’;
return ‘no match’.
```

salient WordNet synsets. By manually selecting correct mappings between the 30 terms from the OSM Semantic Network and WordNet synsets, we obtained a human-generated mapping M_h, which includes 114 correct mappings for the OSM Semantic Network, and 122 mappings for the GeoNames ontology. For the purpose of replication, these test datasets are available online.[12]

4.2 Evaluation measures

To evaluate the performance of *Voc2WordNet*, we define the following performance measures (see Table 1 for notations). Following Euzenat (2007), we assume that a correct mapping belongs to the machine and human mapping $m \in M \wedge m \in M_h$, while an incorrect mapping only belongs to the machine mapping, i.e. $m \in M \wedge \; m \notin M_h$. Hence, we define precision P and recall R of mapping M as:

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

As a general trade-off in the semantic mapping between the OSM Semantic Network and WordNet, we favour precision over recall. In other words, false negative mappings are preferred to false positives. To combine the two measures into a single measure of performance that favours precision over recall, we use a F-measure, defined as:

$$F_{M\beta} = \frac{(1+\beta^2) \cdot P_M \cdot R_M}{\beta^2 P_M + R_M} \quad \beta = .5,\; F \in [0,1] \tag{3}$$

where $\beta = .5$ puts more emphasis on precision than recall. 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 ($M \cap M_h = \emptyset$). This measures are used as indicators of the quality of the semantic mapping in the next sections.

4.3 Preliminary experiments with LIMES

To verify the need for *Voc2WordNet*, we tackled the problem of mapping between a vocabulary and WordNet with existing semantic matching tools. In

[12] See files `osm_semantic_network.manual_wordnet_mapping.rdf` and `geonames.manual_wordnet_mapping.rdf` at `http://github.com/ucd-spatial/OsmSemanticNetwork`

particular, we performed the linkage between the OSM Semantic Network and WordNet with the *LInk discovery framework for MEtric Spaces* (LIMES), described in Section 2.1.[13] Although the Silk framework (Volz et al., 2009) provides similar functionality, LIMES was preferred because of its efficiency and the guarantee of full recall on all the possible mappings.

In order to align the OSM Semantic Network with WordNet, several configurations of LIMES were defined. LIMES computes potential mappings in two given datasets by combining string similarity measures on specific fields. In this context, relevant fields to be compared are the key and value of the OSM concept (*osnp:keyLabel* and *osnp:valueLabel*). In WordNet, the fields are the synsets' definitions (*wns:gloss*) and the corresponding word senses' labels (*rdfs:label*). The string similarity of these four fields can be used to compute the mappings. The fuzzy string similarity function based on *trigrams* was applied to the fields. Pairs obtaining a similarity equal to or greater than a given threshold are included in the mapping.

Using LIMES, we computed the entire mapping between 4,363 OSM concepts and 71,691 WordNet noun synsets using two different strategies, one using only the concepts' labels, and one focused on the lexical definitions. The mappings were then evaluated against the human-generated evaluation dataset, computing precision and recall for each case. When matching OSM concepts and WordNet synsets only based on their labels (e.g. 'amenity=university' and 'university'), the mapping contains very few relevant synsets (max $P_M = .24$, with a similarity threshold $\geq .9$). This experiment also obtained low recall ($R_M < .1$), due to the lack of mappings with related terms from the lexical definitions. As the system has no information about the semantic salience of specific word senses, all the word senses are included.

The other set of experiments was performed on the lexical definitions of the OSM concepts (*skos:definition*) and those of WordNet synsets (*wns:gloss*). In this case, the mapping obtained even lower recall and precision, suggesting that a simple string similarity function applied on definitions does not capture their semantic salience. These two experiments show that, while the basic functionality provided by frameworks such as LIMES is useful in several contexts, especially with very large datasets (Ngomo and Auer, 2011), specific strategies such as *Voc2WordNet* are needed to generate an appropriate mapping between a vocabulary and WordNet. The next section details the evaluation of *Voc2WordNet*.

4.4 Experiment set-up

The algorithm *Voc2WordNet* takes five parameters: V, W, ol_{min}, f_{min}, and Θ (see Section 3). 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:

[13] The experiments were conducted with LIMES v.0.6.

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*

Table 3: Salient synsets in the upper part of the WordNet taxonomy

- Salient taxonomy Θ: either $\Theta \equiv W$ (i.e. taxonomy disabled), or a taxonomy of geographic terms (2 options);
- Minimum lexical overlap ol_{min}: $\{0, 1, 2, \ldots 10\}$ (11 options);
- Minimum word sense frequency f_{min}: $\{0, 1, 2, 3, 4, 5, 10, 20, 30, \ldots 100\}$ (18 options);

These parameters result in $2 \cdot 11 \cdot 18 = 396$ unique combinations of parameters. A random disambiguation approach is added as a baseline. In order to disambiguate the terms from the OSM Semantic Network and the GeoNames ontology to the corresponding word sense in WordNet synsets, we select a subset of the WordNet taxonomy Θ that is relevant to the geographic domain. By manually observing the upper level of WordNet (i.e. synsets with depth $\leq$ 3), we selected eight synsets as roots of the salient taxonomy (see Table 3). 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 was executed on the 396 parameter combinations, parallelised in ten separate threads on both evaluation datasets.

4.5 Experiment results

The experiment generated 396 mappings of the OSM Semantic Network and 396 mappings for the GeoNames ontology. Each mapping was compared with the human-generated dataset described in Section 4.1, obtaining precision, recall, and F-measure. In order to analyse the impact of each parameter on the results, we summarise the performance indicators in Table 4, showing the mean precision $\bar{P}_M$, recall $\bar{R}_M$, and F-measure $\bar{F}_M$. Although *Voc2WordNet* performs better on the OSM Semantic Network ($P = .92, R = .98, F = .92$) than on the GeoNames ontology ($P = .86, R = .9, F = .71$), the results show highly consistent patterns across the two datasets. As expected, precision and recall tend to be inversely proportional. All of the three salience indicators (Θ, f_{min}, ol_{min}) have a positive impact on precision, and negative on recall.

In the case of the OSM Semantic Network, 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. On the GeoNames ontology, the gain in precision is smaller but still detectable. The filter based on f_{min} increases the mean precision at

Parameter	Parameter	*OSM Sem. Net.*			*GeoNames ontology*		
name	value	$\bar{P}$	$\bar{R}$	$\bar{F}$	$\bar{P}$	$\bar{R}$	$\bar{F}$
Salient taxonomy Θ	off	.79	.5*	.67	.77	.40*	.61
	on	.88*	.49	.73*	.79*	.36	.62*
Minimum word	*(off)* 0	.82	.56*	.71	.77	.44*	.62
sense frequency	1	.84	.56*	.72*	.77	.43	.63*
f_{min}	2	.84	.54	.71	.77	.42	.62
	3	.84	.53	.71	.77	.41	.62
			...			...	
	20	.85	.45	.7	.79	.35	.62
	30	.85	.44	.7	.8	.33	.61
	100	.86*	.4	.69	.81*	.32	.61
Minimum lexical	*(off)* 0	.7	.82*	.71	.61	.6*	.59
overlap ol_{min}	1	.75	.81	.75*	.65	.59	.62
	2	.87	.49	.74	.8	.41	.67*
	3	.88	.37	.68	.82	.3	.61
			...			...	
	7	.89	.35	.68	.83	.3	.61
	8	.9*	.35	.68	.84*	.3	.61
Upper bounds	–	.92	.98	.92	.86	.9	.71

Table 4: Summary of experiment results. Mean precision ($\bar{P}$), mean recall ($\bar{R}$), and mean F-score ($\bar{F}$). (*) Best results.

the expense of the mean recall on both datasets, obtaining the best results when $f_{min} = 1$. The minimum lexical overlap ol_{min} has a similar effect on the performance, generating the best results when $ol_{min} = 1$ and 2. These results confirm the validity of the key ideas behind *Voc2WordNet*, described in Section 3.2, indicating that each of the three filters contributes to improve the overall quality of the mapping.

Given that our objective is to maximise the F_M score, biased towards precision, all the three filters need to be utilised in *Voc2WordNet*. In particular, the highest F_M is obtained when the salient taxonomy Θ filter is on, the minimum frequency f_{min} is 1, and the minimum overlap ol_{min} is 1 for the OSM Semantic Network, and 2 for the GeoNames ontology. For the OSM Semantic Network, the selection of these optimal parameters (Θ on, $f_{min} = 1$, $ol_{min} = 1$) results in $P_M = .91$, $R_M = .98$, and therefore $F_M = .92$. For the GeoNames ontology, the best results consist of $P_M = .81$, $R_M = .45$, and $F_M = .7$. These results confirm that *Voc2WordNet* is able to generate a high-quality semantic mapping, vastly outperforming generic tools such as LIMES.

This performance indicates that the *Voc2WordNet* encountered considerably more difficulties with GeoNames terms than with the OSM Semantic Network. By manually inspecting the mappings, it is possible to notice that, compared with the OSM Semantic Network, the GeoNames ontology tends to contain specific and technically complex terms, such as *talus slope*, *salt pond*, *interfluve*, *cuesta*, and *oxbow lake*, which are more challenging to map than common terms such as *mountain* or *road*, resulting in lower precision. Another reason that accounts for the lower recall is the fact that definitions in GeoNames are more concise, with an average of 10.9 words per definition, while the OSM Semantic Network definitions have on average 38.8 words. While OSM definitions are indeed noisier than those in GeoNames, this case highlights that the algorithm suffers from a limited information problem when the lexical definitions are too concise.

A possible solution to mitigate this limitation and increase the recall could consist of extending the search for similar terms in WordNet by visiting related terms. Although performance improvements are certainly possible, as is discussed in the next section, we consider these results satisfactory for the evaluation of our approach to semantic mapping *Voc2WordNet*. The precision, recall, and F-measures obtained by *Voc2WordNet* are comparable with the performance of the state-of-the-art ontology alignment techniques recently evaluated in the context of the Ontology Alignment Evaluation Initiative.[14] The full mapping between the OSM Semantic Network and WordNet, performed with the optimal parameters, is available online as part of the network.

5 Conclusions

Linked open data (LOD) constitutes a promising paradigm to create a shared semantic space, in which heterogeneous geospatial datasets can inter-operate. In the LOD cloud, WordNet can be used as shared semantic ground to enable inter-operability between heterogeneous vocabularies. In this paper, we described our contribution to the LOD vision. First, we outlined a semantic mapping algorithm, *Voc2WordNet*, which aims at generating semantic links between a given vocabulary and WordNet. This algorithm 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. Second, we evaluated *Voc2WordNet* on a random sample of terms from the OSM Semantic Network, and from the GeoNames ontology, obtaining a satisfactory performance.

Voc2WordNet provides a semantic support tool to exploit LOD in geo-applications, increasing the integration of datasets at the schema level. Using WordNet as a semantic hub enables the discovery of implicit semantic relations between features, such as subsumption or meronomy, as well as the discovery of affordances, a promising approach to computational modelling the role of places. Through federated queries over the LOD cloud, these semantic mappings can support tasks at the instance level, facilitating the matching of the same entities across LinkedGeoData, DBpedia, GeoNames, and other geo-knowledge bases (Ballatore et al., 2013).

[14] `http://oaei.ontologymatching.org/2012/results`

Despite the advances reported in this article, our proposal for the bootstrapping of geo-vocabularies in the LOD cloud presents a number of limitations and open challenges. WordNet is a general-purpose semantic resource, and its coverage of geographic terms is limited. While the proposed mapping technique is effective with common terms (e.g. *bay*, *city*, *university*), it would not perform well with many technical terms in highly specialised vocabularies, such as the CORINE Land Cover of the European Environment Agency. As usual in the case of semantic techniques, the generated mappings contains inevitably some degree of noise, ambiguity, and incorrect semantic mappings. SKOS mapping relations are semantically limited, and cannot express the complexity of identity relations discussed by Halpin et al. (2010). Whether a specific semantic mapping is fit-for-purpose, depends on the application in which LOD is being used. For example, a precision of .8 could be sufficient for data exploration, but could be impractical to execute complex spatial reasoning procedures. Future work should include the comparison of other resources as semantic hubs, such as DBpedia and the GeoNames ontology. A larger sample of manual mappings will help evaluate the techniques more thoroughly.

Structuring geographic information 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 (Jain et al., 2010), and spatio-temporal reasoning (Janowicz et al., 2012). These issues notwithstanding, the LOD cloud provides the potential for a vast, open laboratory to a growing community of scientists, software developers, and GIS specialists. Integrating datasets with WordNet is one of the avenues towards the accomplishment of that vision.

Acknowledgements

The research presented in this article 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.

References

Auer, S., J. Lehmann, and S. Hellmann (2009). LinkedGeoData: Adding a Spatial Dimension to the Web of Data. In *Proceedings of the International Semantic Web Conference, ISWC 09*, Volume 5823 of *LNCS*, pp. 731–746. Springer.

Ballatore, A. and M. Bertolotto (2011). Semantically Enriching VGI in Support of Implicit Feedback Analysis. In K. Tanaka, P. Fröhlich, and K.-S. Kim (Eds.), *Proceedings of the Web and Wireless Geographical Information Systems International Symposium*, Volume 6574 of *LNCS*, pp. 78–93. Springer.

Ballatore, A., M. Bertolotto, and D. Wilson (2013a). Geographic Knowledge

Extraction and Semantic Similarity in OpenStreetMap. *Knowledge and Information Systems 37*(1), 61–81.

Ballatore, A., M. Bertolotto, and D. Wilson (2013b). Grounding Linked Open Data in WordNet: The Case of the OSM Semantic Network. In S. Liang, X. Wang, and C. Claramunt (Eds.), *Proceedings of the Web and Wireless Geographical Information Systems International Symposium (W2GIS 2013)*, Volume 7820 of *LNCS*, pp. 1–15. Springer.

Ballatore, A., D. Wilson, and M. Bertolotto (2012). The Similarity Jury: Combining expert judgements on geographic concepts. In S. Castano, P. Vassiliadis, L. Lakshmanan, and M. Lee (Eds.), *Advances in Conceptual Modeling. ER 2012 Workshops (SeCoGIS)*, Volume 7518 of *LNCS*, pp. 231–240. Springer.

Ballatore, A., D. Wilson, and M. Bertolotto (2013). A Survey of Volunteered Open Geo-Knowledge Bases in the Semantic Web. In G. Pasi, G. Bordogna, and L. Jain (Eds.), *Quality Issues in the Management of Web Information*, Volume 50 of *Intelligent Systems Reference Library*, pp. 93–120. Springer.

Berners-Lee, T., J. Hendler, and O. Lassila (2001). The Semantic Web. *Scientific American 284*(5), 28–37.

Bizer, C., T. Heath, and T. Berners-Lee (2009). Linked Data – The Story So Far. *International Journal on Semantic Web and Information Systems. 5*(3), 1–22.

Euzenat, J. (2007). Semantic precision and recall for ontology alignment evaluation. In *Proc. 20th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 348–353.

Euzenat, J., C. Meilicke, H. Stuckenschmidt, P. Shvaiko, and C. Trojahn (2011). Ontology Alignment Evaluation Initiative: six years of experience. In *Journal on data semantics XV*, Volume 6720 of *LNCS*, pp. 158–192. Springer.

Fellbaum, C. (Ed.) (1998). *WordNet: An electronic lexical database.* Cambridge, MA: MIT Press.

Fellbaum, C. (2010). WordNet. In R. Poli, M. Healy, and A. Kameas (Eds.), *Theory and Applications of Ontology: Computer Applications*, pp. 231–243. Springer.

Gangemi, A., N. Guarino, C. Masolo, and A. Oltramari (2003). Sweetening WordNet with DOLCE. *AI magazine 24*(3), 13–24.

Giunchiglia, F., V. Maltese, F. Farazi, and B. Dutta (2010). GeoWordNet: A Resource for Geo-Spatial Applications. In *The Semantic Web: Research and Applications, ESWC 2010*, Volume 6088 of *LNCS*, pp. 121–136. Springer.

Goodwin, J., C. Dolbear, and G. Hart (2008). Geographical Linked Data: The Administrative Geography of Great Britain on the Semantic Web. *Transactions in GIS 12*, 19–30.

Hahn, R., C. Bizer, C. Sahnwaldt, C. Herta, S. Robinson, M. Bürgle, H. Düwiger, and U. Scheel (2010). Faceted Wikipedia Search. In *Business Information Systems*, Volume 47 of *Lecture Notes in Business Information Processing*, pp. 1–11. Springer.

Halpin, H., P. Hayes, J. McCusker, D. McGuinness, and H. Thompson (2010). When owl:sameAs Isnt the Same: An Analysis of Identity in Linked Data. In *The Semantic Web – ISWC 2010*, Number 6496 in LNCS, pp. 305–320. Springer.

Hart, G. and C. Dolbear (2013). *Linked Data: A Geographic Perspective.* Boca Raton, FL: CRC Press.

Isele, R. and C. Bizer (2012). Learning expressive linkage rules using genetic programming. *Proceedings of the VLDB Endowment 5*(11), 1638–1649.

Jain, P., P. Hitzler, P. Yeh, K. Verma, and A. Sheth (2010). Linked Data is Merely More Data. In *AAAI Spring Symposium on Linked Data Meets Artificial Intelligence*, pp. 82–86. AAAI.

Janowicz, K., S. Scheider, T. Pehle, and G. Hart (2012). Geospatial Semantics and Linked Spatiotemporal Data: Past, Present, and Future. *Semantic Web – Special Issue on Linked Spatiotemporal Data and Geo-Ontologies*, 1–13.

Lin, H., J. Davis, and Y. Zhou (2009). An Integrated Approach to Extracting Ontological Structures from Folksonomies. In *The Semantic Web: Research and Applications*, Volume 5554 of *LNCS*, pp. 654–668. Springer.

Mendes, P., M. Jakob, A. García-Silva, and C. Bizer (2011). DBpedia Spotlight: Shedding Light on the Web of Documents. In *Proceedings of the 7th International Conference on Semantic Systems*, pp. 1–8. ACM.

Miles, A., B. Matthews, M. Wilson, and D. Brickley (2005). SKOS Core: Simple Knowledge Organisation for the Web. In *International Conference on Dublin Core and Metadata Applications, DC-2005*, pp. 3–10. DCMI Publications.

Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Computing Surveys 41*(2), 10:1–10:69.

Ngomo, A.-C. N. and S. Auer (2011). LIMES: a time-efficient approach for large-scale link discovery on the web of data. In *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence-Volume Volume Three*, pp. 2312–2317. AAAI Press.

Noy, N. (2004). Semantic Integration: A Survey Of Ontology-Based Approaches. *SIGMOD Record 33*(4), 65–70.

Purves, R. and C. Jones (2011). Geographic Information Retrieval. *SIGSPATIAL Special 3*(2), 2–4.

Ramage, D., A. Rafferty, and C. Manning (2009). Random walks for text semantic similarity. In *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing*, pp. 23–31. ACL.

Scharffe, F., Y. Liu, and C. Zhou (2009). RDF-AI: An Architecture for RDF Datasets Matching, Fusion and Interlink. In *Workshop on Identity, Reference, and Knowledge Representation (IR-KR) at the 21st International Joint Conference on Artificial Intelligence (IJCAI-09).*

Suominen, O. and E. Hyvönen (2012). Improving the Quality of SKOS Vocabularies with Skosify. In *Knowledge Engineering and Knowledge Management*, Volume 7603 of *LNCS*, pp. 383–397. Springer.

Volz, J., C. Bizer, M. Gaedke, and G. Kobilarov (2009). Silk – A Link Discovery Framework for the Web of Data. In *Proceedings of the 2nd Workshop about Linked Open Data on the Web (LDOW2009)*, pp. 559–572.