Crowdsourced geodata and the semantic web: a use case

Gianfranco Gliozzo

University College London, Gower Street, London, WC1E 6BT
Email: g.gliozzo@ucl.ac.uk

KEYWORDS: Neogeography, semantic web, usability, OpenStreetMap, Linked Open Data

1. Introduction

This paper focuses on the opportunities for Geographical Information sharing deriving from Web 2.0 application (O'Reilly 2007). Geographical content created online by users is increasing in quantity, heterogeneity and coverage. This phenomenon is called Neogeography (Turner 2006) or Volunteered Geographical Information (VGI) (Goodchild 2007). OpenStreetMap (OSM) is the most prominent project in existence that gathers the available geodata contributed by users on the web and makes the data freely available to the general public (Haklay & Weber 2008). Understanding the processes for rendering this collected geodata more useful is prominent objective within VGI research (Haklay 2010). The intended benefactors of the online data will be those concerned with day-to-day spatially based decisions and, as such, in need of restricted, localised information regarding amenities and facilities.

Yet, the OpenStreetMap project does not impose strict and obligatory standardised taxonomies. Users are free to participate and add any kind of geographical related information they like. Without strict standardisation, the resulting databases emerging from online communities might suffer from low thematic accuracy owing to synonymic terms (having the same meaning but spelt differently) and homonymic terms (spelled identically but with different meanings). Alternatively, the non-specialised user of the data may not be aware of OpenStreetMap community standardisation practices as they seek spatial information regarding common locations. Similar problems have been encountered during efforts to merge heterogeneous databases. Semantic conflicts arise in these environments when heterogeneous databases containing similar objects have to be merged, hence a "naming conflict" (Naiman & Ouksel 1995) appears caused by the coexistence of homonymic and synonymic terms. Naming conflicts involving synonymy and homonymy are especially problematic to the merging process since part of the information related to a noun is usually embedded in the structure of the database and the application that runs on it. The solution to naming conflicts resides in the semantic description of data and in the use of a linguistic resource able to manage synonymy.

2. Beyond databases

2.1 Bridging Neogeography and linguistics

The ideal environment for the resolution of such conflicts resides in languages capable of expressing semantic relations between data as well as being able to manage synonymy via semantic relations between terms. Ontology has been defined as an explicit and formal specification of a conceptualisation of a domain of interest (Gruber 1993). Ontology therefore embeds the concept of data and meaning related in a domain. The Semantic Web (Berners-Lee et al. 2001) is envisioned as the environment where semantic technologies are used and

knowledge is semantically interconnected. The first large-scale bootstrapping effort towards the semantic web is the web of linked data whose main initiative is the Linking Open Data (LOD) (Bizer et al. 2009a). In the LOD (Figure 1 below) semantic resources from almost every area of human knowledge are published and linked semantically.

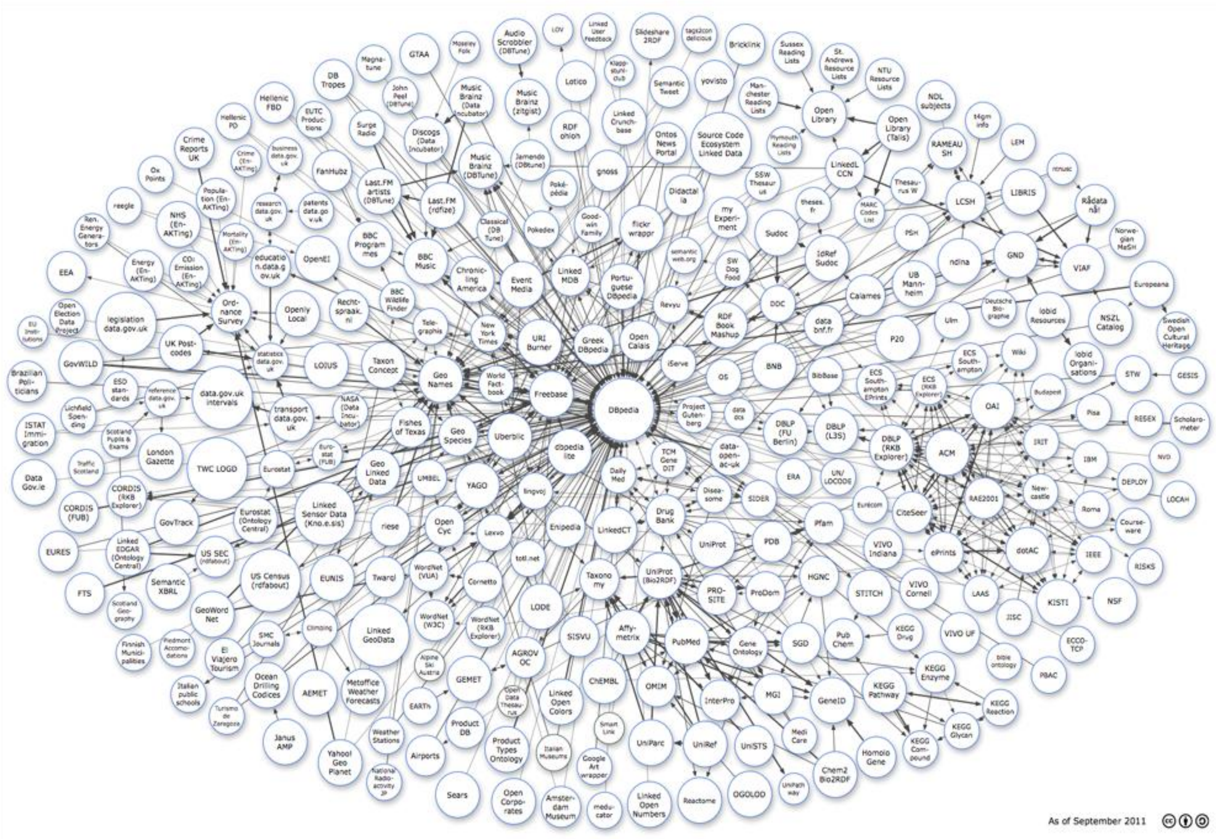

Figure 1. The Linking Open Data cloud as of September 2011

The semantic resources are designed in a way they are both human and machine readable and understandable. In this work the semantic technologies, tools, resources and ongoing initiatives involving OpenStreetMap shall be analysed and compared. Geographic Information research is already looking to semantic technologies for various benefits and solutions and some examples of current research relating to semantic technologies include comparisons between OpenStreetMap and its semantic translation called LinkedGeoData (Auer et al. 2009), and examples of Geographic Information applications involving semantic technologies (Tomai & Kavouras 2004). The present work will show how the flowing of human reasoning through interconnected concepts is mimicked in the semantic web.

To solve the aforementioned naming conflict, semantic technologies and semantic web-based geographic and linguistic resources have to work in unison and we have to deal with the complexities of matching them. The solution is not only pursued using the semantic standardized relations between data; it is deeply rooted in the semantic web itself since it is not intended as a standalone application. There has not been any local copy of data. The work therefore is an example and an evaluation not only of the semantic improvement of OpenStreetMap but also a concrete example of an application running entirely in the web of semantically interlinked data. The creation and testing of semantic web based queries evaluates in a real use case the efficacy and usability of the semantic web resources already published. The flowing of the queries may involve at least three interconnected resources on

the semantic web. In Figure 2, a caption of the LOD cloud diagram and the three semantic resources used in the present work.

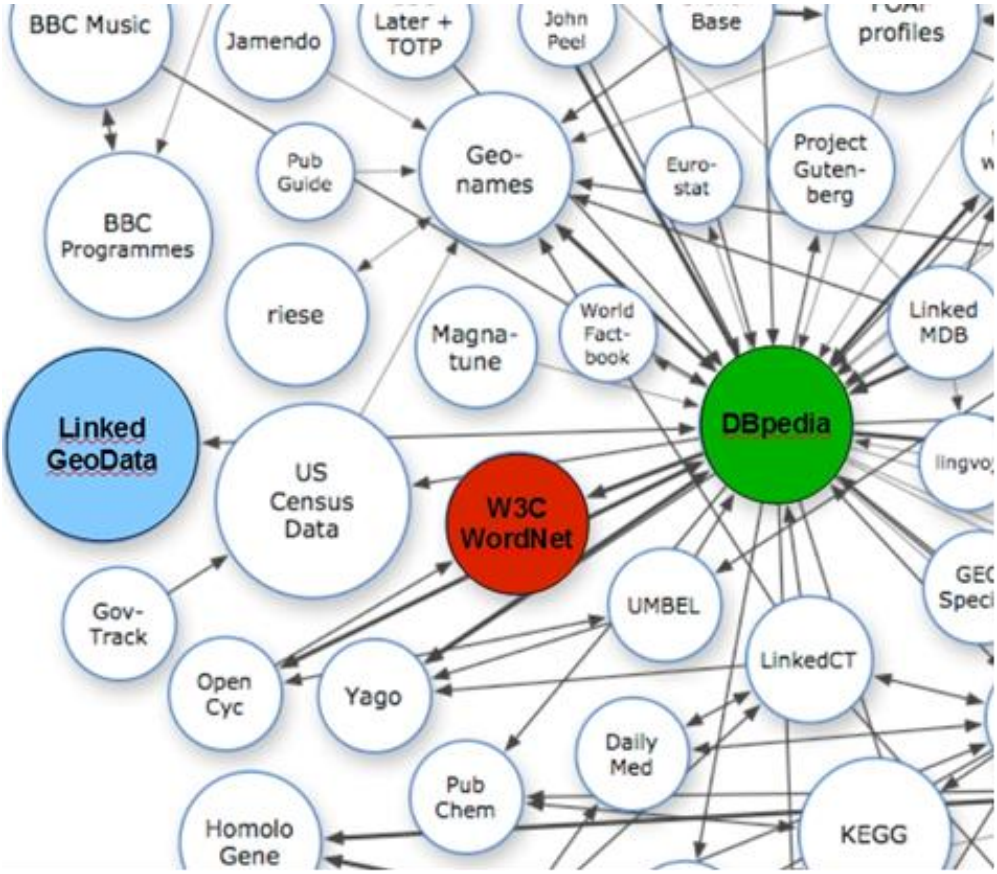

Figure 2. Involved resources on the LOD cloud

The linguistic resource used to solve naming conflicts is WordNet W3C (Assem et al. 2006) the semantic translation of WordNet the prominent linguistic database originally published by Princeton University. WordNet W3C has been translated by teams of experts; it is a semantically built resource. Naming conflict is managed via the creation of a semantic query expansion that searches the web to match a queried term with the relative linguistic semantic resource proceeding then to identify geo objects originally published online by OpenStreetMap users. The linguistic resource in the semantic web has not been related directly to the geographical one. As an intermediate resource, DBpedia the semantic translation of Wikipedia plays a pivotal role (Bizer et al. 2009). DBpedia has been built through the creation of algorithms applied to Wikipedia. DBpedia therefore just like LinkedGeoData is a derived semantic resource. The queries therefore will evaluate the way geographical information is structured in DBpedia. Moreover will be evaluated the way geographical concepts in DBpedia are linked to the two opposite sides of WordNet W3C and LinkedGeoData.

2.2 Methodology

A set of keywords have been selected with two criteria. Most of them have been selected amongst undocumented in OSM but commonly used tags. Some well documented keyword has been selected amongst subclasses of “shop” to assess the local impact of Neogeography and its relevance in the semantic web. The final list of keywords in Table 1, following page.

Table 1. Sample keywords

bakery	frequency	hiking	parking	site
direction	gritting	maintenance	path	stadium
footway	grocery	network	shelter	technology

The creation of two SPARQL (SPARQL Protocol and RDF Query Language) queries tested the reliability of links in the semantic web and the chosen linguistic query expansion. A first query has been based on semantic links (graph patterns) is described by the following Figure3.

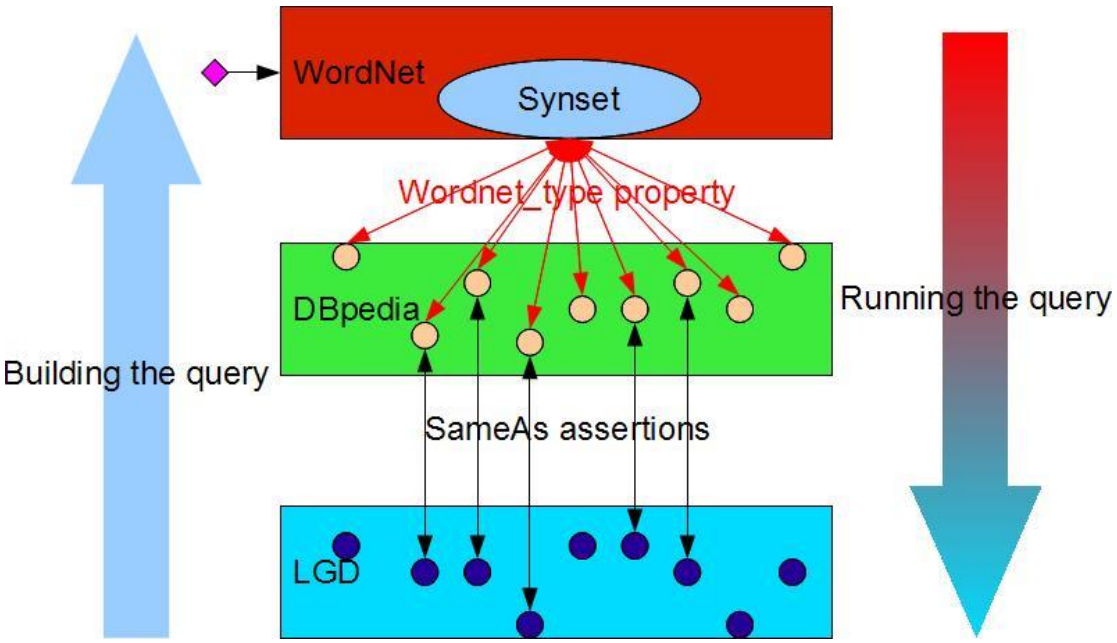

Figure 3. Graph pattern query diagram

The second query has been constructed avoiding the DBpedia ontology since from the first tests the reliability of geographic information taxonomies in DBpedia and their semantic relations with linguistic and neogeographical data resulted fallacious. The second query therefore required a quite more resource demanding research through a string-matching to avoid the DBpedia ontology as depicted in Figure 4. The query is therefore made up by the

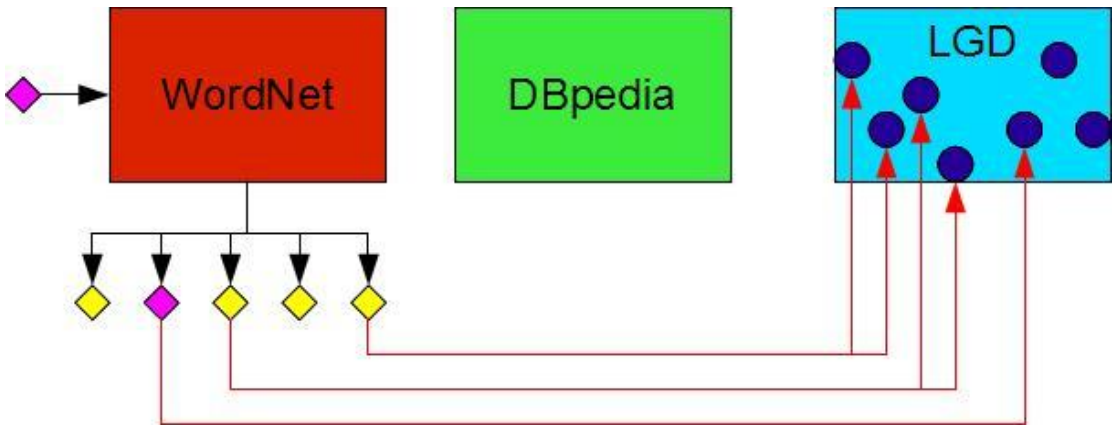

Figure 4. String matching query diagram

linguistic query nested inside the query over the LGD ontology. No semantic pattern has been used to connect the two parts of the query. In the following Listing 1 the SPARQL query following a semantic pattern along the WordNet W3C ontology (lines 8-13) is nested in a string matching query (lines 2-5).

```
1.  PREFIX wn20schema: <http://www.w3.org/2006/03/wn/wn20/schema/>
2.  select ?geo ?vocab
3.  from <http://linkedgeodata.org/> where
4.  {?geo a ?vocab
5.  FILTER regex(str(?vocab), ?lexforms)
6.  {SELECT distinct ?lexforms
7.  from <http://wordnet.rkbexplorer.com/>
8.  WHERE {?bWords wn20schema:lexicalForm ?lexforms .
9.  ?bWordSense wn20schema:word ?bWords.
10. ?aSynset wn20schema:containsWordSense ?bWordSense .
11. ?aSynset wn20schema:containsWordSense ?aWordSense .
12. ?aWordSense wn20schema:word ?aWord .
13. ?aWord wn20schema:lexicalForm "bakery".}}}
```

Listing 1. The String matching query

3. Results and future work

In the following Table 2 the results of the two queries over the LOD are compared with simple queries to the two main ontologies involved, LikedGeoData and WordNet.

Table 2. Comparing results

Terms			**Objects in LGD**		
Sample keywords	**Synonyms in WordNet**	**Used synonyms in LGD**	**Direct LGD query for the keyword**	**Graph pattern query1**	**String matching query2**
grocery	5	2	192	0	41.694
path	6	2	95.596	0	6.058.107
parking	1	1	191.325	0	191.325
shelter	3	1	4.395	0	4.390
footway	1	1	669.487	0	0
hiking	1	1	1.418	0	0
technology	4	1	29	0	29
frequency	5	0	0	0	0
vending	1	1	2.927	0	2.926
agricultural	3	2	7	0	7
snowmobile	1	1	4	0	4
bakery	3	2	7.239	0	7.238
network	7	2	2	0	201
gritting	0	0	0	0	0
stadium	4	3	5.244	310	5.262

As anticipated the graph pattern query gave very poor results working only for one keyword over fifteen. The graph pattern query was very effective in terms of execution time (less than

a minute). The second query performed very well but every query lasted three to four hours. The table should be read as follows: the first block made up by three columns is dealing with synonymy amongst terms; the second block provides an outline of the linguistic query expansion grounded in LGD terms. The first column is the list of the keywords chosen amongst several tags in OSM. The second column reports how many synonymous terms can be found starting from the keywords of the first column querying the WordNet W3C ontology. A zero in this column means that the keyword is not in WordNet W3C. In the third column there is the number of synonyms of the keyword that are used in the LGD ontology. As an example the keyword "stadium" has 4 entries in WordNet (stadium, sports stadium, bowl, arena) but only three of them are used in LGD (stadium, bowl, arena). In the right hand side of the table there are three columns to evaluate the effectiveness of the query expansions. The forth column gives the number of spatial objects that you get querying directly the LGD ontology using the keywords in the first column. The fifth column gives the result achieved applying the first query, the graph pattern, as shown in Figure 3. The sixth column gives the result achieved applying the string matching query.

The challenge has been the exploration of viable paths along the three heterogeneous resources. The semantic resources have been built using different methodologies and conceived and developed by different communities. Two facets are carried out through this research: the short-term aspect, where the query expansion is developed and evaluated; and the broader perspective, accumulating all the evaluations performed for the integration of geographical user generated content with semantic web technologies and resources. This work underlines how a core component to manage geographical information from heterogeneous origins can be implemented in the semantic web. The so diverse performance of the two queries as reported on Table 2 demonstrate the need of further developments in the management of crowdsourced geographical information on the semantic web. The need to design a second approach and the way it has been performed derives from a deeper analysis of the semantic resources involved in the queries. Such analysis underlined the fact that semantically related information and real world entities representation in the semantic web provide a very powerful opportunity that has been poorly implemented. The heterogeneity of origins will be a permanent feature of the semantic web since not only finely analysed and structured information sources are published in the LOD but also resources coming from online communities. It implies that the LOD embodies also data loosely structured both logically and semantically (Specia & Motta 2007). As a further demonstration of the above mentioned weaknesses there is a most recent approach to the semantic translation of the OSM database that has been developed and natively linked with WordNet and DBpedia (Ballatore et. al 2012). The present work has also evaluated the meeting point of two heterogeneous communities; on one side those of OpenStreetMap and Wikipedia prone to volunteer in data creation and on the other side the scientists that developed in a quite structured and task oriented way WordNet W3C.

Performing the present work a series of evaluations and suggestions have been achieved, they cover several issues. The work demonstrates how technological resources and publication procedures for the semantic web have to be fine-tuned. The structure of geographical information in the semantic web and the pivotal role that has been given to DBpedia has to be considered thoroughly. The way semantic information is elicited from crowdsourced taxonomies and transformed in ontologies (therefore called folksonomies) and the intermediate steps between the elicitation and the publication on the semantic web have to be considered carefully.

6. Acknowledgements
This paper is based on research that was carried out within the MSc GIMA programme. I am grateful for their support to Aldo Gangemi from the Semantic Laboratory Lab (STLab) of the Italian National Research Council (CNR) in Rome, Italy, and Rob Lemmens from the Faculty of Geo-Information Science and Earth Observation of the University of Twente, ITC in Enschede, The Netherlands.

References
Assem M, Gangemi A, Schreiber G, 2006, RDF/OWL Representation of WordNet W3C Working Draft 19 June 2006.
Auer S, Lehmann, J, & Hellmann, S, 2009. LinkedGeoData – Adding a Spatial Dimension to the Web of Data. *The Semantic WebISWC 2009*, 5823, 731–746. doi:10.1007/978-3-642-04930-9_46 Springer Berlin Heidelberg.
Ballatore, A., Bertolotto, M., & Wilson, D. C. 2012. Geographic Knowledge Extraction and Semantic Similarity in OpenStreetMap. *Knowledge and Information Systems*, 1–21 Springer Berlin Heidelberg.
Berners-Lee T, Hendler J, Lassila O, and others. 2001. The semantic web. *Scientific American* 284, no. 5: 28–37.
Bizer C, Heath T, Berners-Lee T, 2009a, Linked Data - The Story So Far. *International Journal on Semantic Web & Information Systems*, Vol. 5, Issue 3 Information Resources Management Association ITJ5383
Bizer C, Lehmann J, Kobilarova G, Auer S, Becker C, Cyganiak R, Hellmann S, 2009. DBpedia – A Crystallization Point for the Web of Data. *Web semantics*, Volume: 7, Issue: 3 (September 2009), pp: 154-165
Goodchild M F, 2007, *Citizens as sensors: the world of volunteered geography*. GeoJournal 69, no. 4: 211–221.
Gruber T. R., 1993, A translation approach to portable ontologies. *Knowledge Acquisition.* Vol. 5, No. 2,
Haklay M, 2010, How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets, *Environment and Planning B: Planning and Design* 37(4) 682 – 703
Haklay M, Weber P, 2008. OpenStreetMap: user-generated street maps. *IEEE Pervasive Computing* 7, no. 4: 12–18.
Naiman C F, Ouksel A M, 1995, A classification of semantic conflicts in heterogeneous database systems. *Journal of Organizational Computing and Electronic Commerce* 5, n°. 2: 167–193.
O'Reilly T, 2007, *What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software* (30 September 2005). Available at http://oreilly.com/web2/archive/what-is-web-20.html [Accessed 15 November 2012].
Specia L, Motta E, 2007. Integrating folksonomies with the semantic web. *The semantic web: research and applications*: 624–639.
Tomai, E., and M. Kavouras. 2004. From "onto-geonoesis" to "onto-genesis": The design of geographic ontologies. *Geoinformatica* 8, no. 3: 285–302.
Turner, A. 2006. Introduction to Neogeography. O'Reilly.

Biography
Gianfranco Gliozzo is a current PhD student at University College London. Gianfranco holds a Master Degree in Construction Engineering and an MSc in GIS. He is interested in spatial planning, geography, linked data. His current research interests involve the support of biodiversity and conservation initiatives through citizen science.