

# GIMA

Geographical Information Management and Applications

## Bridging the gap between user generated spatial content and the semantic web

**Master of Science Thesis**

**Gianfranco Gliozzo**

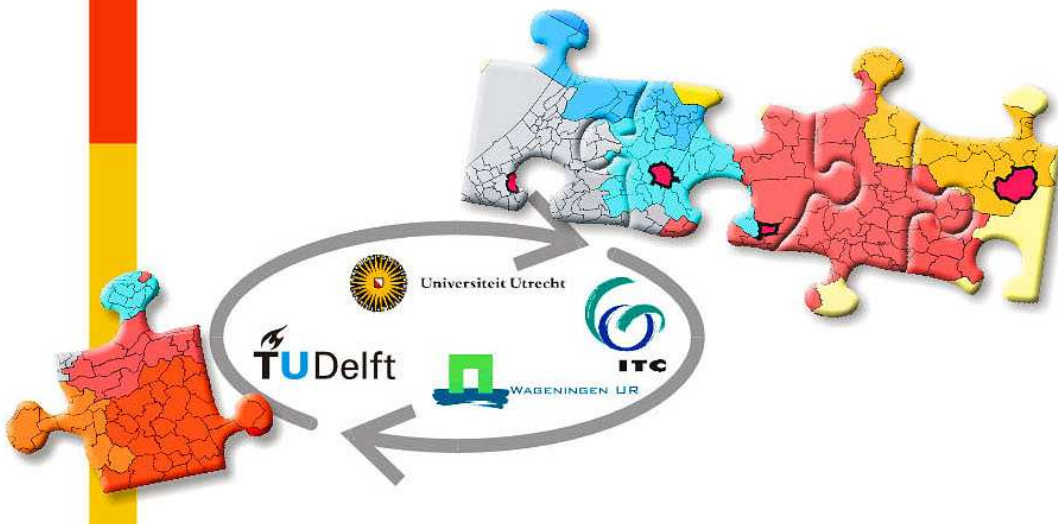
December 10<sup>th</sup> 2010

Professor: Prof. Dr. M.J. (Menno-Jan) Kraak,  
Department of Geoinformation Processing  
Faculty of Geoinformation Science and Earth Observation  
University of Twente

Supervisor: Dr. Ir. Rob Lemmens  
Department of Geoinformation Processing  
Faculty of Geo-Information Science and Earth Observation  
University of Twente

Supervisor: Aldo Gangemi  
Senior Researcher  
Semantic Technology Lab (STLab)  
Institute for Cognitive Science and Technology,  
Italian National Research Council (ISTC-CNR)

Reviewer: Mw. drs. M.E. (Marian) de Vries  
OTB Research Institute for the Built Environment  
Delft University of technology



Gianfranco Gliozzo Msc GIMA Master thesis  
Bridging the gap between user generated spatial content and the semantic web

## Acknowledgments

The present work could not have been completed without the loving support of my parents and my sister, old and new friends made on the occasion of the thesis. Let me quote my cousin, Alfio, always ready to stimulate and to suggest visions from a different perspective, he is the “guilty”, who suggested me the way of the semantic web and introduced me to Aldo Gangemi.

Aldo Gangemi I will always be grateful for his generous support.

The fantastic community of OpenStreetMap with in particular I want to thank simone (Simone Cortesi) for our long and pleasant conversations on OpenStreetMap story and perspectives and for his kind suggestions, tosky as well (Luigi Toscano) and David Paleino for their generous and tireless support in the struggle with open source applications and operating systems.

## Table of contents

Acknowledgments .....	iii
Table of contents.....	iv
Table of figures .....	vii
Table of tables.....	viii
Abstract.....	1
PART I: TECHNOLOGIES, TOOLS AND RESOURCES .....	3
1    Chapter one: background. ....	4
1.1    Introduction .....	4
1.2    Web2.0 or social web .....	4
1.3    Geoweb2.0.....	5
1.4    Quality of Crowdsourced data.....	5
1.5    VGI data quality analysis .....	6
1.6    Modelling the problem: semantic heterogeneity in multiple databases .....	7
1.7    Naming conflict: solving approaches .....	7
1.8    Web 3.0 or semantic web.....	8
1.9    Geoweb3.0.....	9
1.10    A step further – innovation aimed at .....	10
Research aims .....	10
1.11    Problem setting .....	10
1.12    Problem statement .....	12
1.13    Research objectives and questions.....	12
1.14    Research Limitations .....	13
Extent limitations: .....	13
1.15    Research Method .....	14
Method Extent limitations.....	14
Research Depth limitations .....	14
1.16    Thesis development .....	15
2    Chapter two: The web environment: from web 2.0 to the semantic web. ....	16
Introduction.....	16
2.1    The web of data .....	16
2.2    Semantic web stack .....	16
2.3    Queries in SPARQL .....	19
2.3.1    Query forms .....	19
2.3.2    Dealing with precision .....	20
2.4    Ontologies .....	20
2.4.1    Formalised ontologies .....	21
2.4.2    Ontology matching.....	22
2.4.3    Elicited ontologies .....	22
2.4.4    Ontologic precision/expressivity .....	23
2.4.5    Ontologies and folksonomies.....	24
2.5    The environment of the proposed development: Linking Open Data.....	25
2.5.1    DBpedia .....	26
2.5.2    Geodata in DBpedia.....	26

2.6	WordNet and WordNet RDF/OWL .....	27
2.6.1	WordNet data model .....	28
2.6.2	Previous applications of WordNet in GI .....	29
2.6.3	WordNet in the semantic web .....	29
2.7	Folksonomies and the semantic web .....	30
2.8	Maturity of ontology development .....	32
3	Chapter three: Geoweb 2.0 and OpenStreetMap .....	33
	Introduction .....	33
3.1	OpenStreetMap and VGI .....	33
3.2	Data in OpenStreetMap .....	36
3.3	The OSM structure .....	37
3.4	Data format .....	38
3.5	Database .....	39
3.6	Tools to extract data .....	42
4	Chapter four: OpenStreetMap and the semantic web .....	43
	Introduction .....	43
4.1	Geographic Information ontologies .....	43
4.2	Semantic efforts and OpenStreetMap .....	44
4.3	The Semantic MediaWiki – Machine readable map feature list effort .....	45
4.4	LinkedGeoData .....	46
4.4.1	Overview .....	46
4.4.2	The LinkedGeoData ontology .....	46
4.4.3	Ontology evaluation .....	49
4.4.4	LinkedGeoData mapping with DBpedia .....	50
4.4.5	LinkedGeoData and OpenStreetMap - Taxonomies .....	52
4.4.6	LinkedGeoData and OpenStreetMap - Data comparison .....	53
4.5	SVG maps .....	55
4.6	OpenStreetMap Wrapper .....	55
4.7	Evaluation of semantic initiatives for OpenStreetMap .....	56
	PART II: THE DEVELOPMENT .....	58
5	Chapter five: Improving OpenStreetMap retrieval using semantic technologies .....	59
	Introduction .....	59
5.1	Use case, constraints and advantages of the chosen approach .....	60
5.1.1	The constraints .....	60
5.1.2	The advantages .....	61
5.1.3	The procedure .....	61
5.2	The sample .....	63
5.3	Matching semantic resources on the LOD .....	66
5.4	Query1 - The graph pattern .....	67
5.4.1	LinkedGeoData mappings with DBpedia .....	69
5.4.2	DBpedia mappings with WordNet RDF/OWL .....	71
5.4.3	Extracting synonyms in WordNet RDF/OWL .....	72
5.5	Query1 Graph pattern runtime .....	74
5.5.1	Analysis of the results of query1 graph pattern .....	77
5.6	Query 2 - String matching .....	78
5.6.1	Query construction .....	79
5.6.2	Testing the query .....	81

5.6.3	Query 2 String matching runtime.....	84
5.6.4	Analysis of the results of query 2 string matching .....	85
5.7	Overall reflections on chapter five results.....	87
5.7.1	Query results comparison .....	87
5.7.2	The core component long lasting development .....	88
5.7.3	Using the core component for an application .....	89
6	Chapter six: Conclusions and recommendations. ....	91
	Introduction.....	91
6.1	The main objective of the thesis.....	91
6.2	Reflection .....	92
6.3	Main results of the thesis.....	93
6.3.1	Main conclusions .....	93
6.3.2	Answering the research questions.....	94
6.4	Guidelines to choose between the two procedures to query building .....	98
6.5	Recommendations and Further developments .....	99
6.5.1	Recommendations for LinkedGeoData development.....	99
6.5.2	Further developments - Extending the core component with existing published resources .....	101
6.5.3	Extending the core component in the future: Geo SPARQL .....	101
7	References .....	104
8	Appendix: LinkedGeoData type of data with a mapped instance with DBpedia ..	115

## Table of figures

<i>Figure: 1-1 The evolution of web and related technologies. From (Berners-Lee, Hendler, 2001).</i> .....	9
<i>Figure: 2-1 The Semantic Web Technology Stack taken from (Bratt, 2006) .....</i>	17
<i>Figure: 2-2 Layered ontologies according to Guarino (1998).....</i>	21
<i>Figure: 2-3 The Linking Open Data cloud as of September 2010.....</i>	25
<i>Figure: 2-4 Places in DBpedia ontology.....</i>	27
<i>Figure: 2-5 WordNet RDF/OWL class hierarchy. Using RDF Gravity.....</i>	30
<i>Figure: 3-1 The growth of OpenStreetMap - users and points .....</i>	34
<i>Figure: 3-2 Google Maps and OSM comparison - De Uithof.....</i>	35
<i>Figure: 3-3 Google Maps and OSM comparison - Amsterdam Zoo. ....</i>	35
<i>Figure: 3-4 Google Maps and OSM comparison – Cyprus.....</i>	36
<i>Figure: 3-5 OSM components and parts.....</i>	37
<i>Figure: 4-1 LinkedGeoData database, from (Auer, et al. 2009, 1).....</i>	47
<i>Figure: 4-2 LinkedGeoData upper level hierarchy.....</i>	48
<i>Figure: 5-1 Involved resources on the LOD cloud.....</i>	66
<i>Figure: 5-2 Query1 - Graph pattern blocks.....</i>	68
<i>Figure: 5-3 Query1 Graph pattern at instance level.....</i>	69
<i>Figure: 5-4 Extracting synonyms using WordNet RDF/OWL.....</i>	72
<i>Figure: 5-5 Query2 - String matching blocks.....</i>	79
<i>Figure: 5-6 Query2 String matching, listing explanation.....</i>	80
<i>Figure: 5-7 Evolution of resources towards the core component. ....</i>	89
<i>Figure: 5-8 Application envisaged.....</i>	90

## Table of tables

Table: 1.1 Tagwatch statistics Netherlands as of December 2009.....	11
Table: 3.1 OSM tags of nodes as currently rendered.....	40
Table: 3.2 Metadata and location of current nodes in OSM.....	40
Table: 3.3 OSM ways through nodes sequence.....	40
Table: 3.4 OSM tags for ways.....	40
Table: 3.5 OSM ways metadata.....	41
Table: 3.6 OSM relation members.....	41
Table: 3.7 OSM relation tags.....	41
Table: 3.8 OSM relations metadata.....	41
Table: 4.1 Geographic information ontology elements.....	44
Table: 4.2 Geographic information ontology elements and LinkedGeoData.....	50
Table: 4.3 Evaluation of ontologies designed for OpenStreetMap.....	56
Table: 5.1 Sample keywords from national statistics.....	63
Table: 5.2 - Sample keywords.....	64
Table: 5.3 Sample keywords: types, subclasses superclasses.....	65
Table: 5.4 Sample keywords and mapping between LinkedGeoData and DBpedia.....	70
Table: 5.5 Stages and variable values of bakery synonym extraction using WordNet RDF/OWL.....	73
Table: 5.6 LinkedGeoData types of data from mapped "stadium".....	77
Table: 5.7 Sample keywords synonyms.....	82
Table: 5.8 Synonyms of sample keywords and LinkedGeoData instances.....	83
Table: 5.9 Query 2- string match results.....	84
Table: 5.10 Evaluation of query2- string match.....	86
Table: 5.11 Overall evaluation of queries results.....	87
Table: 6.1 SQLMM functions in Openlink Virtuoso.....	103



## List of abbreviations

Abox	Assertion Box
CC BY SA	Creative Commons Attribution-Share Alike 2.0 Generic
DOLCE	Descriptive Ontology for Linguistic and Cognitive Engineering
GFM	General Feature Model
GML	Geography Markup Language
GPS	Global Positioning System
GUI	Graphical User Interface
HTML	Hyper Text Markup Language
HTTP	Hyper Text Transfer Protocol
IAOA	International Association for Ontology and its Applications
LGD	LinkedGeoData
LOD	Linked Open Data
OAEI	Ontology Alignment Evaluation Initiative
OGC	Open Geospatial Consortium
OSM	OpenStreetMap
OWL	Web Ontology Language
POI	Point Of Interest
RDBMS	Relational Data Base Management System
RDF	Resource Description Framework
RDFS	Resource Description Framework Schema
SDI	Spatial Data Infrastructure
SMW	Semantic Media Wiki
SPARQL	SPARQL Protocol and RDF Query Language
SVG	Scalable Vector Graphics
SW	Semantic web
Tbox	Terminology Box
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
VGI	Volunteered Geographic Information
W3C	World Wide Web Consortium
WMS	Web Mapping Service
XML	eXtensible Markup Language
XSLT	eXtensible Style sheet Language Transformations

## Abstract

This work concerns the development opportunities for Geographical Information deriving from new technologies as well as web applications in use today. Geographical content created online by users is increasing in quantity and heterogeneity. OpenStreetMap is the most prominent project in existence that gathers the available geodata contributed by users on the web. Understanding the processes for rendering this collected geodata more useful is the objective of this thesis. 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.

Without strict standardization, the resulting databases emerging from online communities will be afflicted with low thematic accuracy owing to synonymic terms. Alternatively, the non-specialised consumer may not be aware of OSM community standardization efforts 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. *Naming conflicts* involving synonymy and homonymy are especially problematic to the merging process. The solution to naming conflicts resides in the semantic description of data and in the use of a linguistic resource able to manage synonymy.

The ideal environment for the resolution of such conflicts therefore resides in languages capable of expressing semantic relations between data as well as being able to manage synonymy via semantic relations between nouns. This environment is the Semantic Web. In the semantic web, semantic resources from almost every area of human knowledge are published and linked semantically. In this thesis 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 and Geographic Information are included herein, these include comparisons between OpenStreetMap and its semantic translation called LinkedGeoData, and between LinkedGeoData and examples of Geographic Information applications involving semantic technologies. 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 linguistic resource used to solve naming conflicts is the semantic translation of WordNet, the prominent linguistic database originally published by Princeton University.

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

Gianfranco Gliozzo Msc GIMA Master thesis  
Bridging the gap between user generated spatial content and the semantic web

online by OpenStreetMap users.

Two facets therefore underlie this thesis: 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 to explore the possibility of further information source integration.

Gianfranco Gliozzo Msc GIMA Master thesis  
Bridging the gap between user generated spatial content and the semantic web

## **PART I: TECHNOLOGIES, TOOLS AND RESOURCES**

# **1 Chapter one: background.**

## **1.1 Introduction**

In the last decade, the evolution of the internet into Web 2.0 has brought with it an enormous diffusion of online content and tools. Users can add content by tagging online resources or adding features or points of interest. Consequently, the management of Geographic Information published online has overcome the confines of GIS technicalities, opening itself to a broader arena with the so-called Geoweb2.0. What once was the niche of Geographic Information specialists is now a field where almost anyone can use background maps for their own applications through APIs and mashups. Geographic Information content is now also collected through users' tags and local knowledge is increasing in quantity and diversity. You can find through web search engines almost any data built by users with location on maps. Most of the initiatives that create geodata online lack strict standardization in order to empower users and foster their participation. Some of the geographical initiatives on the web are creating data that can be downloaded, but problems arise in the usability of this data. The quality of peer-produced geodata can be subject to various deficiencies. In this thesis, the thematic (also referred to as 'attribute' or 'semantic') accuracy will be examined. Data collected publicly without strict standardization introduces cultural deviations such as polysemy and synonymy corresponding in the degradation of the quality and usability of resulting geo databases. Every user inserts data according to his personal background using its associated tags. Where standardization is missing, it can be assumed that every user is creating his own database. Given this scenario, this thesis will discuss the methodologies used to overcome similar issues in merging heterogeneous databases. Semantic technologies can be the means to resolve such problems, they are crucial in the development of the Semantic web or web 3.0, and thus resolving semantic problems in data via semantic technologies can provide the basis for new possibilities. Semantic content is based on data stored online that can be matched and made to work together in a human and machine interpretable fashion. Using open geodata online and semantic technologies will not allow only the previously outlined linguistic limitations to be overcome but, due to the semantic data residing in the web, geodata will be able to be integrated into any kind of human knowledge resource and be human and machine interpretable.

## **1.2 Web2.0 or social web**

The web 2.0 (O'Reilly, 2007) is an improved web, driven by fiddling users

(Unwin, 2005) who submit new data to the web by means of pictures, comments, tags, GPS observations, every kind of user generated content (Scharl & Tochtermann, 2007). This is a kind of social/bottom up web where user-generated content in a business environment is named *crowdsourcing* (Howe, 2006) and the world of peer production is referred to as *wikinomics* (Tapscott & Williams, 2006). The principles of web 2.0 applied in the GI field brought about Geoweb2.0.

### **1.3 Geoweb2.0**

The more commonly used web 2.0 applications are Wikipedia, del.icio.us, Flickr, Facebook and Twitter; while in the geographical environment OpenStreetMap, Google maps, Wikimapia, Geocommons, MapuFature and Flickr can be cited. They are shaping a new approach to geographical information. The phenomenon of geographical applications in the web 2.0 has been labelled as volunteered geographic information (VGI) (Goodchild, 2007a) and Neogeography (Turner, 2006) (Haklay, Singleton, Parker, 2008) or geoinformation bottom up (Bishr & Kuhn, 2007) in distinction to PPGIS (Public Participation GIS) (Tulloch, 2008). Everyone with internet access can be a “sensor” thus facilitating the creation and the updating of geographical information. (Goodchild, 2007a+b).

Crowdsourcing relies on the users or social group background (Elwood, 2008b) it is a kind of citizen science (Goodchild, 2008). The rising of a network of trusted users can foster the governance of the projects towards an improvement of the usability of VGI dataset (Bishr & Kuhn, 2007). Governance of web 2.0 projects is provided through the creation of a community amongst people (users) that have developed the project. Thus, real time geographical information can have a crucial impact on emergency response as recently exemplified by the Haiti earthquake where online mappers from all over the world contributed in photo interpreting aerial pictures of Port Au Prince to find places where refugees were meeting and camping and to indicate broken bridges or collapsed buildings. Almost 169 contributors were helping rescue workers within hours by mapping Port Au Prince from post earthquake aerial pictures.

### **1.4 Quality of Crowdsourced data**

Geoweb 2.0 is a great opportunity but some weaknesses are shared with all the web 2.0 applications. The above mentioned redundancy of applications can result in the production of different datasets with spots of information (Haklay, 2008). Quality issues also arise from web 2.0 practices. Some VGI initiatives are totally missing meta-information on tags thus hindering the usability of collected data (Bishr & Kuhn, 2007) (Castelli et al., 2007).

Datasets are created by different users applying different accuracy and production times that are not often explicitly known. (Maué & Schade, 2008).

Most of the time, contributors are not professional geographers or mappers thus resulting in inconsistent information (Bishr & Mantelas, 2008) (Goodchild, 2008) (Flanagin & Metzger, 2008).

VGIs are favoured amongst all web 2.0 applications in overcoming some of these weaknesses. Once displayed in maps, Geographic information to some extent provides a good initial overview of the above-mentioned weaknesses. Due to the willingness to promote participation, successful VGI datasets have limited standardization allowing more specific and comprehensive coverage of local areas and culture, where “global” counterparts for example special restaurants for some kind of local food do not exist.

## 1.5 VGI data quality analysis

The quality and usability of VGI datasets are the subject of debate, valuable works (Haklay, 2008) (Coote & Rackham, 2008) underlined the different elements of spatial data quality as stated in (ISO/TC 211, 2002), where internal and external quality measures are identified (Devillers & Jeansoulin, 2005). For external qualities, accuracy and precision for VGI datasets have been measured and evaluated using GIS tools (Aather, 2009) (Haklay, 2008). Amongst internal quality measures, thematic accuracy is still a doubtful issue for VGI datasets (Coote & Rackham 2008). Often in Geographic Information literature and practice, thematic accuracy is only seen as the optimal interpretation of ground values coming from remote sensing; in fact, it is often measured using only the misclassification matrix as in Goodchild (1995). Moreover in GI practice thematic inaccuracy is mitigated *“through clear definition in capture specifications, operator training, and logic-based quality control measures built into data capture and editing systems”* (Harding, 2005): all elements that are missing in VGI initiatives. Therefore, since VGI databases rely on every single user background (Elwood, 2008b) (Goodchild, 2008) the possibility that users use synonymic terms to tag similar entities is very high. Moreover, since tagging until now has only been in English, it relies also on single user awareness of English<sup>1</sup> language semantics and environment. It may include, for instance, certain problems associated with false friends<sup>2</sup> in European languages. The particular nature of VGI that relies on single user cultural background cannot be verified or corrected easily. In fact Coote & Rackham (2008), commenting on OpenStreetMap data quality, wrote, *“This would not appear to be an approach conducive to achieving a high level of thematic accuracy or very helpful to users”*, since every user inserts data according to his personal background using their own logical tags.

---

<sup>1</sup> Even the distinction between British and American English will affect queries results.

<sup>2</sup> [http://en.wikipedia.org/wiki/False\\_friend](http://en.wikipedia.org/wiki/False_friend)

## **1.6 Modelling the problem: semantic heterogeneity in multiple databases**

In VGI initiatives where standardization is missing, you can assume that every user is creating his own database with its own internal thematic accuracy. The problem then, can be seen through the perspective of the creation of a unique database (the global database) coming from users' own databases (the local databases). In databases, we can distinguish between the schema or intention of a database and data values, instances or extensions of the database. In our case, there is no need for schema integration but only integration at the data value level. The problem of synonymic terms in local databases to be merged is the "naming conflict" according to Naiman and Ouksel (1995).

## **1.7 Naming conflict: solving approaches**

A problem similar to the one we encountered in crowdsourced data has been analyzed and dealt with in recent decades by several works coming from the database environment. In 1986, there were already several methodologies (Batini et al., 1986) to perform heterogeneous database integration. The main issue has always remained the integration between different database schemas since, in the cited paper, only two methodologies tackle the semantic naming conflict at the instance level. The solution proposed for synonymy at the instance level relies on the creation of a "cross reference lexicon" of names or in the human choice between proposed pairs of objects with a high degree of similarity that have been coupled automatically by the integration application. Over the years, the main focus has remained on schema integration. Research over this period has suggested solving the naming conflict between database values using external references. The creation of an intermediate global schema to convey all semantic knowledge stored in the different databases is suggested in (Reddy et al., 1994) while the solution of naming conflict is always left to data dictionaries and human interaction. Li and Clifton developed database integration extracting semantics of databases through neutral networks (Li & Clifton, 1994) (Li et al., 2000) leaving, since 1994, the solving of synonymy and homonymy issues to the creation of a synonym lexicon and in 2000 they specifically suggested using the linguistic resource WordNet (Miller, 1995) or CYC (Lenat, 1995). However, human intervention is still necessary in databases where abbreviations or "compressed" sentences are used instead of simple lexical forms. In the GI field, heterogeneous database integration is perceived as an effort to obtain interoperable datasets or to manage multiple scale datasets. Bishr (1998), to overcome semantic issues creates proxy classes and content that represent semantic only and not geometric relations, between geographic features. Bishr then creates the semantic beyond the data aggregating concepts and elements based on their meaning. Reddy et. Al (1994) in a broader perspective structured their approach similarly. The most recent



developments in heterogonous database integration are moving towards the use of ontology (Hakimpour, & Geppert. 2005) adding semantic description of database schema as in (Lohar, 2010) for the merging of loosely coupled federated databases. Ontology has been defined as an explicit and formal specification of a conceptualisation of a domain of interest (Gruber, 1993). In the GI field, heterogeneous database integration adopted the ontology-based approach as in (Kavouras & Kokla, 2000) where, to solve the naming heterogeneity, they proposed the decomposition of every concept in the semantic construction of its meaning, thus referring to a higher semantic level for solving naming conflicts. In another case, the naming conflict is based on a semantic representation of the knowledge involved. A more knowledge-aware system might be used to overcome the underlined limitation of VGI datasets. In this thesis, an attempt is made to implement a system to harvest and interpret user generated content using semantic technologies. The semantic description of knowledge beyond names is the key to solving naming conflicts as the semantic description of databases is the key to resolving heterogeneous database interaction, integration or merging.

Almost all issues involved in heterogeneous database integration efforts are somehow semantic since information stored in databases is more than merely ordered data. The trend to integrate applications and data either online or locally has driven the development of the semantic technologies that are the backbone of the web 3.0. Web 3.0 or semantic web is, following from web 2.0, a further step in the development of the web that is gaining more importance.

## 1.8 Web 3.0 or semantic web

The forthcoming development of the web is the web 3.0 or semantic web (in the following also indicated as SW) (Berners-Lee et al., 2001) (Gangemi & Mika, 2003) which enables the web to resemble more human reasoning and to improve information sharing (Stuckenschmidt & Van Harmelen, 2004). Comparing it to the current *eyeball web* (Sheth & Meersmann, 2002), which is a web of hyper textually interlinked documents whose meaning is only humanly understandable, the semantic web is a web of data where the meaning of relations between data can be also machine interpreted through ontology, a *software agent web*. Ontology then opens broader perspectives than the ones offered by databases. Reusability and interoperability of knowledge embedded in ongoing applications required a huge effort in expressing information semantically coming from the database schemas. Applications were primarily not developed with reusability in mind (Spyn et al., 2002), database schema, database instances and applications were tightly designed to achieve a specific goal. Data was not initially independent from applications but through ontology, semantic independence of knowledge has been achieved (Meersmann, 2001). In database applications, knowledge is also embedded in the application code that led to the creation of the database (Zhao, & Chang, 2007). Knowledge can be extracted from ongoing applications based on databases (Zhao & Chang, 2007) and conveyed to

ontologies. We began, therefore, with a web of documents connected through standardized protocols (HTTP-HTML), through to a web of standardized data format (XML-RDF) and have now progressed to the field of logically connected data whose main tools are standardized languages able to express semantics of data through ontologies. Data with semantic relationships are then named knowledge bases instead of databases. The above mentioned development is expressed by the evolution of standards and technologies as in the following Figure 0-1, taken from an article published in "nature" (Berners-Lee, Hendler, 2001).

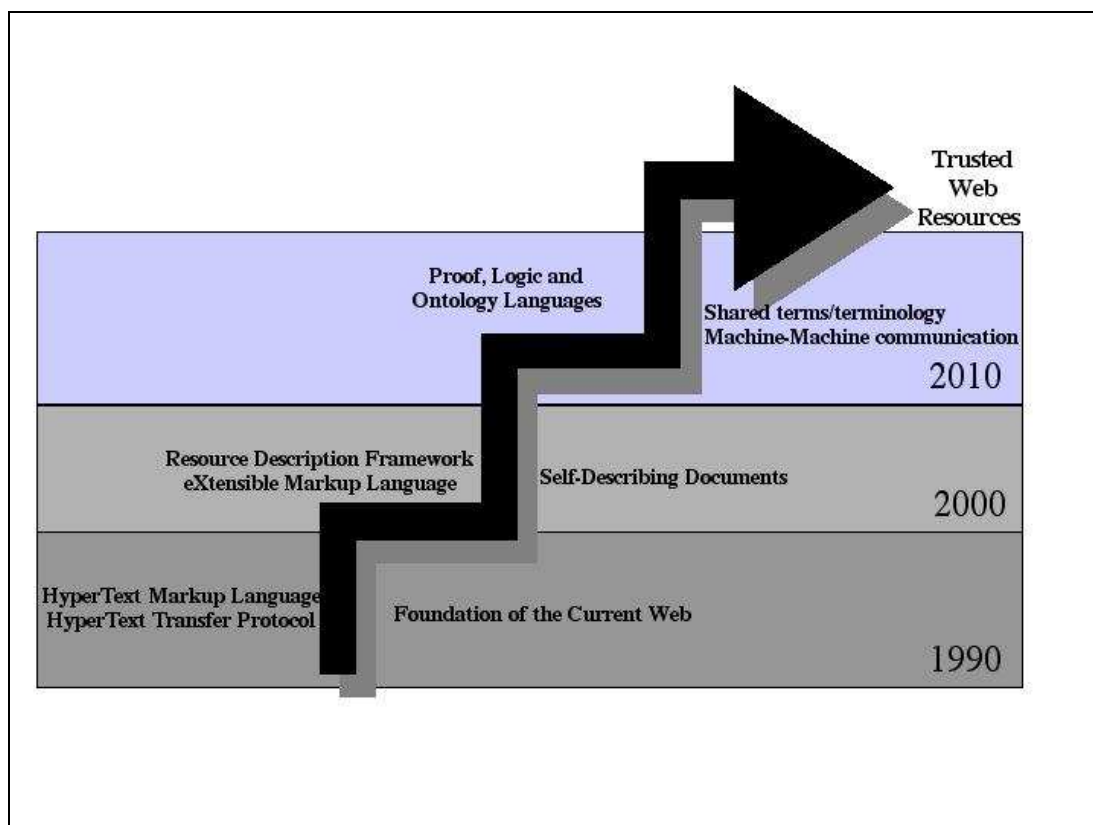


Figure: 1-1 The evolution of web and related technologies. From (Berners-Lee, Hendler, 2001).

## 1.9 Geoweb3.0

The semantics of geodata are important in understanding the meaning of (geographic) information and can be captured in ontologies. In the GI field, several attempts have been carried out on using semantic technologies: (Egenhofer, 2002) (Kuhn, 2005) to enhance interoperability, (Lieberman, 2006) (Lemmens, 2006) (Harvey et al., 1999) (Fallahi et al., 2008) (Fonseca et al., 2002), services coupling (Lemmens et al., 2006), metadata description (Schuurman & Leszczynski, 2006), or to increase the interaction between GI and semantic technologies improving ontology creation (Fonseca, Davis,

Câmara, 2003).

The semantic description of geoinformation can help users by submitting correct and consistent spatial web content integrating different sources of information on the client side, thus improving comprehensiveness (Fonseca et al., 2002).

Semantically well defined and formalised map content and user interfaces that present formalised this content correctly can also lead to enhanced spatial catalogue services (Craglia, 2007) (Goodchild, 2007b) and improve their interaction with the end user. Furthermore, once ontologies for some VGI applications are known, semantic mapping helps to discover similarities in those information structures for the purpose of geoinformation source integration. Moreover, semantically well defined content can be mapped easily to other kinds of information in the semantic web (Stuckenschmidt & Van Harmelen, 2004).

## **1.10 A step further – innovation aimed at**

This thesis attempts to investigate and develop the integration potential between geoinformation and the semantic web. The present work is aimed at the design of a core component for applications willing to use geospatial data obtained through web 2.0 initiatives. On one hand, low thematic accuracy can be overcome in case standardization efforts are not working well and on the other, the generic application user might not be aware of standard terms adopted by online communities thus querying locations with non-standard terms. The most interesting factor is that since semantic web embeds semantics about data geoinformation it will be integrated with other sources of knowledge and results shall be obtained based on knowledge already present in the semantic web, avoiding any programming. In this thesis, results similar to previous research projects on query expansions have been obtained performing the task relying directly on the web of data. It is the starting point on the way to broader integration with other information sources, to more complex ways of using geoinformation in the semantic web.

## **Research aims**

### **1.11 Problem setting**

The focus of this thesis comes from the freedom of tagging that users enjoy mainly in OpenStreetMap (following OSM), the most consistent VGI database. In OSM information on real world objects is inserted through tags in the form of a couple 'key=value' where the key is a term and the value is a specialization or further clarification of the key. (E.g. sport=football or building=yes). One of the mottos of the project is "*Map for the database don't map for the renderers*" and this means that a lot of data inserted in the

database is not rendered through general purpose renderers. Furthermore, many key/value pairs in the map-features<sup>3</sup> list lack a defined render style in the most commonly used styles: maplink and osmarender. Therefore, the database is considerably richer than what the maps show. To favour fiddling mappers, there are no restrictions on tags that can be used in OSM. This means that even the map features list is only a recommendation and not a compulsory repository of tag values. Mappers can therefore use any tags they like but users are required to document them on the OpenStreetMap wiki, even if self explanatory. This openness can drive to a plethora of different tags as documented in *tagwatch*<sup>4</sup>.

As an example of the richness of the database, the following *Table: 1.1 Tagwatch statistics Netherlands as of December 2009* gives a statistic on the tags used in the Netherlands<sup>5</sup>. In the second column, we have the total amount of keys and tags (often used instead of value) and relations used in The Netherlands. The third column extracts the numbers regarding elements that are described in the OSM wiki from the numbers listed in the second column.

Element	Different types	Mentioned the OSM Wiki	in En Description	En Translation
<b>Key</b>	890	157	111	111
<b>Tags</b>	7643	429	426	426
<b>Relations</b>	23	4	4	4

**Table: 1.1 Tagwatch statistics Netherlands as of December 2009**

In the fourth and fifth columns, we have the number of documented elements that have descriptions and translations in OSM's wiki.

Then only 17% of used keys and 5% of used values are mentioned in the OSM wiki, therefore few users are inserting their own tags in the wiki as cited above. It could, albeit surprisingly, wind up that all those different values and tags used without proper coordination are consistent. Another problem is that most of the undocumented tags come from codes used in massively imported datasets. The community or some company working in close contact with the project developed quite a lot of tools to prevent inconsistent editing of which some are automated. Most of them focus on geometric and few in attributes debugging.

Summing up, OpenStreetMap is a broad project in the process of creating an enormous database. Standardization has the form of community agreements and the prevailing sense of community ensures progress.

<sup>3</sup> [http://wiki.openstreetmap.org/wiki/Map\\_Features](http://wiki.openstreetmap.org/wiki/Map_Features)

<sup>4</sup> Statistics on tag usage <http://wiki.openstreetmap.org/wiki/Tagwatch>. Tagwatch is an online service that documents the usage of tags in OpenStreetMap

<sup>5</sup> <http://tagwatch.stoecker.eu/Netherlands/En/index.html> accessed 23 November 2009

## 1.12 Problem statement

Usability of crowdsourced data such as VGI databases created without strict guidance can be undermined by inconsistency from the linguistic point of view and conditions the results of queries posted to such databases. On one hand taggers of web 2.0 applications are mainly spurred on by visibility; (Elwood, 2008b) on the other hand, to foster participation, web 2.0 applications do not require tags to be contextualized and are consequently unstructured pieces of information covering a wide range of fields relating to the web 2.0 application purpose. The above cited geoweb 2.0 applications are mainly focused on the description of places users encounter in everyday life and so all aspects of human knowledge are included in an enormous unstructured and possibly inconsistent collection of data. Usability of data collected through users' tags is thus a troublesome issue mainly for geoweb 2.0 applications.

## 1.13 Research objectives and questions

The initial general objective of the thesis has been the **Enhancement of a VGI dataset**. During the development of the thesis, greater insights have determined certain refinements of the research objectives. Considering all insights shown in the previous sections, the main research objective of the thesis is expressed by the following statement:

***Enhancement of the usability of a VGI dataset achieved overcoming the low thematic accuracy using semantic technologies in the environment of the semantic web.***

It implies a general research question:

Has a way been found to improve the usability of volunteered geographic information?

Going in details in order to achieve the above outlined research objective four sub objectives has been found. The publication on the LinkingOpenData initiative of the semantic translation of the most prominent VGI initiative (OpenStreetMap) gave us the opportunity to outline more operational research objectives.

The first sub objective:

1. *Identify and characterize a VGI initiative*

To achieve the outlined sub-objective the following research questions were posed:

1.1. Why OpenStreetMap?

1.2. Is there any abstraction level or hierarchical structure in the OSM

database?

The second sub objective:

2. *Identify an enhancement for the VGI initiative*

To achieve the outlined sub-objective the following research question was posed:

- 2.1. What kind of improvement can be achieved with the present thesis?

The third sub objective:

3. *Explore the relation between the chosen VGI initiative and the semantic technologies*

To achieve the outlined sub-objective the following research questions were posed:

- 3.1. Why semantic technologies?
- 3.2. Has the broad OSM community attempted semantic developments of the project?
- 3.3. How to deal with ontologies coming from ongoing web 2.0 application?

The fourth sub objective:

4. *Explore the potential of the semantic web for geoinformation*

To achieve the outlined sub-objective the following research questions were posed:

- 4.1. Are the available resources in the Semantic web able to support the task of the present work?
- 4.2. Which way can other information sources in the semantic web be coupled with geoinformation?
- 4.3. Which way to choose if too many tags for the same object lead to enriched or confused data? (How to deal with tag inconsistencies?)
- 4.4. In what terms does the present work remain valid if we apply it to a different geoweb3.0 ontology?

## 1.14 Research Limitations

### **Extent limitations:**

This research will not embark on all the above cited aspects that come after

OSM ontology has been set. We will not try to integrate OSM with other VGI applications through semantic mapping. We will not implement a semantic based SDI. We will not create a GUI to help user tagging. Moreover, we will not implement a migration from OSM databases to semantic knowledge bases.

The linguistic enhancement will be focused only on one aspect of linguistic relations like synonymy. Developments and improvements of existing resources will be identified and suggested but not implemented.

## **1.15 Research Method**

### **Method Extent limitations**

The present work will be developed focusing on usability of existing semantic web tools standards and formalizations. Programming will be avoided.

### **Research Depth limitations**

As within many organizations, the introduction of new ideas in OSM has to be considered carefully. Cultural and organizational needs have to be taken into account (Reeve & Petch, 1999) (de Man & van den Toorn, 2002). Working on the organizational aspects connected with the present work is beyond the objectives of the present work. Only the following considerations led us to change partially the initial plan. According to well designed GI new ideas in organizational environments, the development of OSM might be appropriate for users and the purpose and have to be accepted by the "hosting" company having support both at the highest level of hierarchy and at the user level. Due to the organizational environment of OSM, big changes are difficult to implement, an example is the never ending debate on the license change that is poorly supported by the base of users.

The project is led by a foundation<sup>6</sup> whose members are enthusiastically engaged in the development of the still developing project. It leads to a status quo attitude to maintain the mapping and editing activity as free and easy as it is now<sup>7</sup>. OSM is successful and the challenges foreseen by the foundation board shall have no little or no impact on the development of the work presented in this thesis<sup>8</sup>. Moreover, some issues that the foundation board consider important are not supported by the majority of users like the above cited license change. Therefore, the presented approach focuses more on a semantic post processing activity to improve OSM data usability. The tagging activity will be maintained and kept the way it is by OSM members. In OSM wiki pages, it is often stated that their openness drove them to omit giving strict guidance to tagging. Tagging is voluntary, new tags can be proposed and then evaluated but everyone can use his own. The present work will not try to structure the tagging activity differently.

---

<sup>6</sup> [http://www.osmfoundation.org/wiki/Main\\_Page](http://www.osmfoundation.org/wiki/Main_Page)

<sup>7</sup> <http://lists.openstreetmap.org/pipermail/dev/2008-December/013247.html>

<sup>8</sup> [http://wiki.openstreetmap.org/wiki/Things\\_To\\_Do](http://wiki.openstreetmap.org/wiki/Things_To_Do)

## 1.16 Thesis development

This thesis will develop from an initial introduction to new technologies and terminologies as described in Chapter Two: the web environment from web 2.0 to the semantic web. Here, the main characteristics of the upcoming web 3.0 or 'semantic web' and its sub component 'web of data', will be explained in further detail.

The focus will then progress to peer production of geographic content in Chapter Three: Geoweb 2.0 and OpenStreetMap, where the main features of Volunteered Geographic Information will be explained. OpenStreetMap, the most successful initiative will also be described, especially its parts more related to the development of the thesis. The existing point of contact between OpenStreetMap and the semantic web and the state of this relationship will be investigated in Chapter four: OpenStreetMap and the semantic web. Later, attempts to use semantic technologies to support OpenStreetMap will be shown.

The most recent contributions are developed in chapter five: Improving OpenStreetMap retrieval using semantic technologies. Web 3.0 will be used to improve the usability of GI information created according to VGI principles.

Closing in chapter six: *Conclusions and recommendations*, all the questions that drove the development of the present work will be answered; the results analyzed and some suggestions for further developments will be forwarded.



## 2 Chapter two: The web environment: from web 2.0 to the semantic web.

### Introduction

In this chapter, the technological environment chosen for the thesis development is explained. The semantic web is described in sections 2.1 to 2.4, touching on the technological and methodological issues that will be encountered in the following chapters. In the second part, from section 2.5 to 2.8, the main resources of the semantic web to be matched in pursuance of the research objectives will be described.

### 2.1 The web of data

The semantic web is fostered by the World Wide Web Consortium (W3C)<sup>9</sup> international community, developing web standards to achieve the web's full potential. Here is a description of the semantic web from its inventor (Berners-Lee et al. 2001)

*"The Semantic Web is a web of data (...) The Semantic Web is about two things. It is about common formats for integration and combination of data drawn from diverse sources, where on the original Web mainly concentrated on the interchange of documents. It is also about language for recording how the data relates to real world objects. That allows a person, or a machine, to start off in one database, and then move through an unending set of databases which are connected not by wires but by being about the same thing."<sup>10</sup>*

The semantic web therefore gives us the possibility to relate real word objects to web descriptions created by different users because it renders the web more similar to human reasoning with corresponding improvements in information sharing (Stuckenschmidt & Van Harmelen, 2004). It means that by browsing in the semantic web we can connect geoinformation with any kind of information available in the web of data. In the semantic web environment, an object is called instance and instantiation is the process of populating a class of objects.

### 2.2 Semantic web stack

Since 2001 Berners Lee (2001), forecasting the development of the semantic

---

<sup>9</sup> <http://www.w3.org/Consortium/>

<sup>10</sup> <http://www.w3.org/2001/sw/>

web, outlined the role of a series of technologies and standards required to achieve it. Data resources and relationships between them are identified by Uniform Resource Identifiers (URIs) they are simply Uniform Resource Identifiers of data and relationships between data while Uniform Resource Locators (URLs) refer to documents. The complete structure of the semantic web and the technologies involved are often represented by the Semantic Web Technology Stack developed initially by Tim Berners Lee. A recent representation of it is given in the following Figure 2-1:

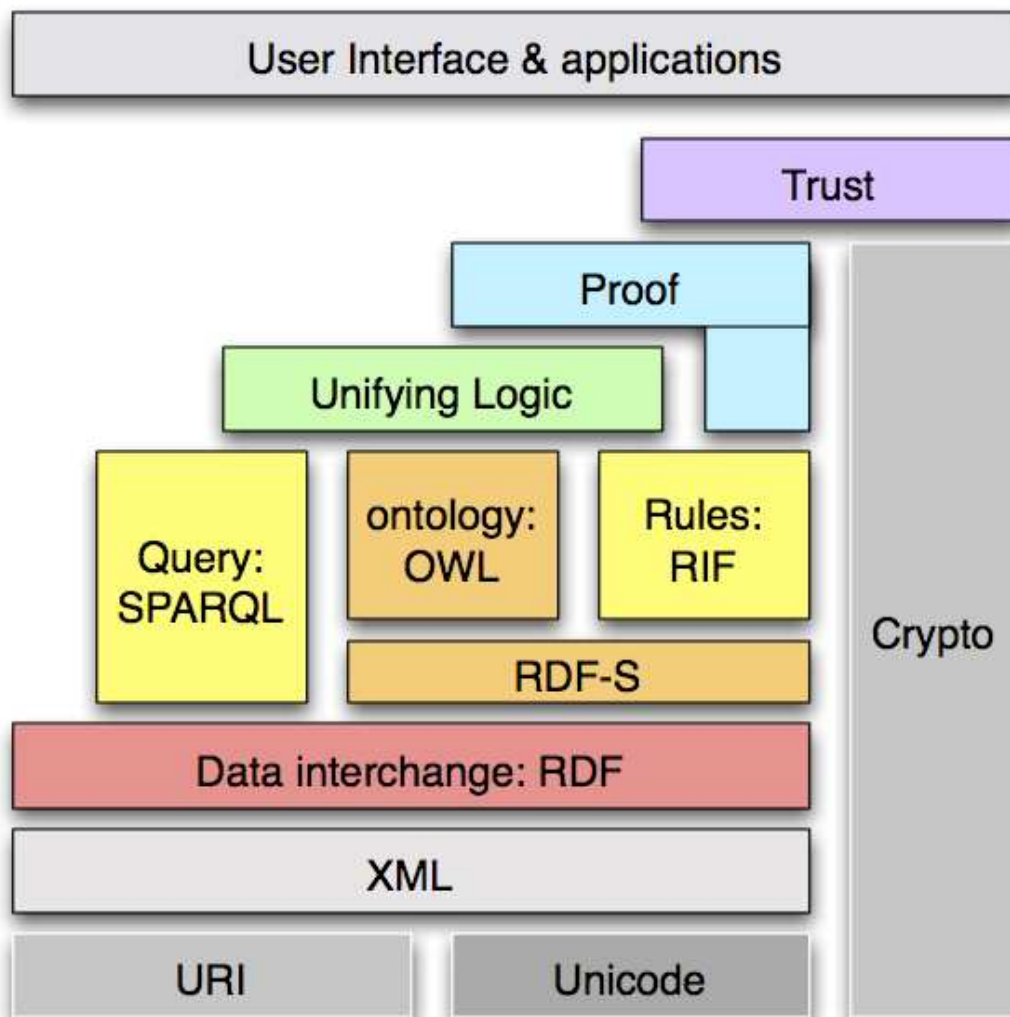


Figure: 2-1 The Semantic Web Technology Stack taken from (Bratt, 2006)

The building blocks of the semantic web are based primarily on the eXtensible Markup Language (XML) (Bray et al., 2000) and the Resource Description Framework (RDF) (Lassila et al. 1999). Through the XML format, data is described in a standardized way. The Resource Description Framework (RDF) expresses basic statements about anything, giving a model and a layered

structure to model assumptions. RDF defines then directed graphs through data and properties.

RDF works through a set of triples with each triple being rather like the subject, verb and object of an elementary sentence. The three elements are separated by a space and every triple ends with a dot. Given S P O. is a triple where S is the subject P is the predicate O is the object. Triples are written using XML syntax. What makes the semantic web an outstanding development of the web is that it is able not only to describe but also to integrate knowledge because subject, predicate and object are *identified by a Universal Resource Identifier (URI) as in a link on a Web page. (URLs, Uniform Resource Locators, are the most common type of URI.) The verbs are also identified by URIs, which enable anyone to define a new concept, or verb, by simply defining a URI for it somewhere on the Web.* (Berners-Lee et al., 2001).

The third main component of the Semantic web is the Web Ontology Language (OWL), a W3C standard issued in 2003. It reached the 2.0 version in 2009 (W3C OWL Working Group, 2009). OWL allows us to accurately express the relationship between classes, entities and properties adding semantic expression to RDF. Different levels of expressivity are also included in RDF, RDFS and in OWL. OWL sublanguages are OWL Lite, OWL DL and OWL Full.

OWL Lite allows for the expression of a simple classification and simple constraints. It performs well and can be easily used for the automated migration of catalogues and thesauri.

OWL DL allows the expression of Description Logics. It opens the maximum degree of expressivity without losing computability and decidability of reasoning systems since all elements can be computed in finite time.

OWL Full provides the maximum level of expressivity but does not guarantee computability. It mainly entails statements about classes and properties.

The equivalence between classes is an important aspect that distinguishes OWL DL and OWL Full. In OWL DL, classes are simply collections of objects and cannot be treated as individuals while in OWL Full all individual constructs can be applied to classes. It is quite an important difference since developers of ontologies prefer the retention of full computability of OWL DL over expressivity as shall be shown in following sections. Through OWL, for instance, different sources of information relating to the same entity can be related stating an equivalence that in the semantic web environment is called mapping. Mappings are implemented using sameAs assertions. To state that A is B we have the following triple:

A <<http://www.w3.org/2002/07/owl#sameAs>> B.

The sameAs assertion allows the expression of full equivalence between individuals in OWL DL, while in OWL Full it can also be used to manage equivalence between classes. Due to the importance of full computability allowed by OWL DL it is the most commonly used OWL sublanguage. It

means that during this work we will have to manage equivalence between classes of objects starting from the instance level.

The fourth component is SPARQL (a recursive acronym, *SPARQL Protocol and RDF Query Language*) the W3C standard RDF query language and protocol (Prud'hommeaux & Seaborne, 2008) that has reached the draft version 1.1 (Harris, Seaborne, 2010). It works on all kind of triples; with version 1.1 nested queries are explicitly allowed. SPARQL is used to query RDF/OWL knowledge bases that are available through SPARQL endpoints on the web.

SPARQL queries are conveyed in several ways. The SPARQL protocol (Clark et al. 2008) SPARQL clients and processors communicate through HTTP bindings. Since every document in the semantic web is accessible through a HTTP, to shorten and to make documents more readable it is possible to put prefixes (also called namespaces in XML) at the beginning of SPARQL queries as well.

Below are some of the most commonly used:

rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

rdfs: <http://www.w3.org/2000/01/rdf-schema#>

owl: <http://www.w3.org/2002/07/owl#>

xsd: <http://www.w3.org/2001/XMLSchema#>

fn: <http://www.w3.org/2005/xpath-functions#>

In ontologies and knowledge bases, it can be useful to distinguish between the Abox and the Tbox. While they are modelled in the same way we collect in the Tbox (terminology box) all the assertions regarding the relationships between concepts, hence class hierarchies of the specified domain of interest. The Tbox focuses then on concepts and classes. In an Abox (assertion box), we collect all relationships between individuals and classes. An Abox is focused on instances, features in the GI environment. This distinction, often not cited in GI literature, will be useful for us.

## 2.3 Queries in SPARQL

Since SPARQL is the main tool to extract information from knowledge bases, its workings are explained in the following sections. SPARQL queries are posted to SPARQL endpoints

### 2.3.1 Query forms

The query forms are the clauses SELECT, CONSTRUCT, ASK and DESCRIBE. When in an RDF triple we have variables, than the triple is called a triple pattern. A set of triple patterns is called graph pattern. Through SELECT we can ask for a variable or more variables in a triple or graph pattern and the result gives the values of the data in the queried knowledge base that matches the triple or graph pattern. For example, if S and P are

known URIs and we want to determine the object in the knowledge base associated with them we will define the ?o variable and our query will have the following form:

```
SELECT ?o WHERE {S P ?o .}
```

We can define more variables and insert a graph pattern into the query. Using CONSTRUCT we ask, using an RDF graph (a graph template), where there is a variable and we gain the queried triples in an RDF graph pattern with not only the value of the variable but a graph with values substituting variables.

Using ASK we can test if a query has answers.

Using DESCRIBE we can identify all resources in a graph that are related to a specific variable.

### 2.3.2 Dealing with precision

Several commands combined can shorten the timing of query results and make easier and less time consuming queries on the enormous web of data (sometimes a query can take hours). The ASK query form serves this purpose. Precision in queries is crucial for usable answers. To enhance precision we have filters on literals, types and numeric constraints as well as limiting the number of answers. The more we know about the data we are looking for, the more easily and quickly it can be targeted. If we know in which graph we can find the information we are looking for, we can direct our query engine to it. The FROM clause is designed to focus on a graph. If we know that the data we are looking for is in a specified RDF dataset and we can identify it with an URI like <http://www.graph.com/examples> we can direct our query to it as in the following:

```
SELECT ?o FROM <http://www.graph.com/examples> WHERE {S P ?o .}
```

If case the RDF dataset contains a named graph it can also be invoked through the FROM NAMED clause.

## 2.4 Ontologies

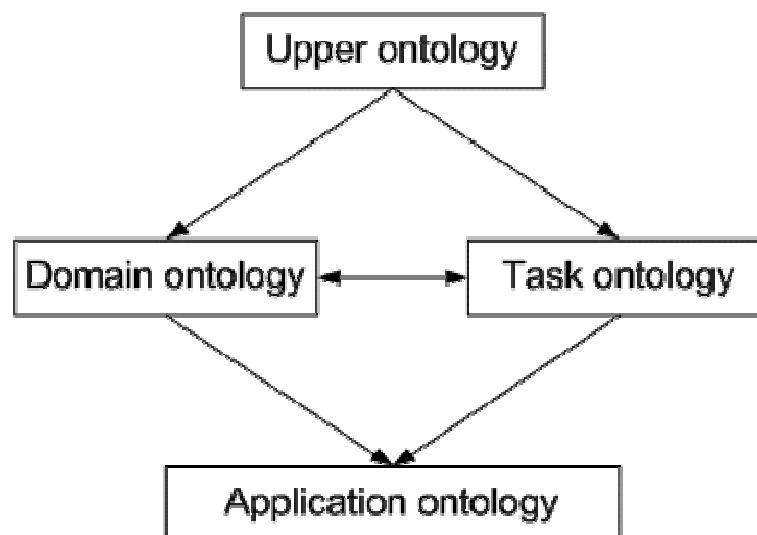
The semantic web is based on ontologies that have been defined as an explicit and formal specification of a conceptualisation of a domain of interest (Gruber, 1993). A more comprehensive definition for ontologies can be found in (Guarino, 1998).

*“An ontology is a logical theory accounting for the intended meaning of a formal vocabulary, i.e. its ontological commitment to a particular conceptualization of the world. The intended models of a logical language using such a vocabulary are constrained by its ontological commitment. An ontology indirectly reflects this commitment (and the underlying conceptualization) by approximating these intended models.”*

In this thesis, ontologies are classified along two axes, in the first they are divided into two classes according to their generating constituents. So formal ontologies (Guarino, 1998) developed by experts pertain to the first class while elicited, bottom up, or social semantics (Gruber, 2007) derived by folksonomies pertain to the second class. The role of the active constituent or actor is therefore incorporated in the definition of ontology as suggested by (Mika, 2007). Along a second axis, knowledge representation methods have been classified according to their ontological precision that starts from the simplest catalogue to the full axiomatized theory (Obrst, 2008) (Smith & Welty, 2001).

#### 2.4.1 Formalised ontologies

Ontologies for information systems are structured in a layered manner (Guarino, 1998) as in *Figure: 2-2 Layered ontologies according to Guarino (1998)*. The layered approach differentiates ontologies into four groups, subdivided into three levels. Upper or top level ontologies or foundational ontologies comprise very general assumptions; they contain very abstract concepts like time space and matter, Upper or top level ontologies are developed by philosophers and knowledge engineers. One example is the DOLCE ontology (Gangemi, et al. 2002.) (Masolo, et al., 2003) used in (Fallahi et al, 2008) in the GI field. Sometimes the lexical WordNet RDF/OWL ontology has been used as an upper level ontology (Fonseca, et al. 2002). Upper ontologies are very rigid and stable in time and are designed in such a way as to contain conceptualizations shared by a wide range of domains.



*Figure: 2-2 Layered ontologies according to Guarino (1998)*

Domain and task ontologies are specific to the domain of interest the ontology deals with. To be more precise, domain ontologies describe the vocabulary related to a particular domain of knowledge. Task ontologies are a specialization of terms described in upper level ontologies with a particular

application or task in mind. This level is also characterized by high stability. Domain and task ontologies are developed by domain experts and are often the translation of standards or conventions for the specific domain. In the GI, filed domain ontologies may be formalizations of OGC and ISO standards as the GFM (OGC, 2009) or GML (OGC, 2003). Examples in GI can be found in (Lemmens, 2006) (Klien & Probst, 2005) (Schade & Maué, 2008). Thus underlining the distinction between real world objects and their GI representations (Tomai & Kavouras, 2004). This is the level where interoperability issues are modelled.

Application ontologies are a further step between general terms and the real world. They set out the particular roles played by domain entities while performing an activity.

The design of formal ontologies have a pivotal role in the semantic web, design patterns have been developed to nurture ontology design (Gangemi 2005) (Gangemi & Presutti, 2009). Information coming from legacy systems can also be processed and interpreted to gain knowledge that databases are not able to infer (Gangemi et al. 2007).

#### **2.4.2 Ontology matching**

Ontology matching is one of the most disputed arguments in the SW. It consists in the matching between ontologies coming from different environments. There are several initiatives and research projects still in progress. The ontology alignment initiative (OAEI)<sup>11</sup> is testing ontology matching tools on different use cases. The OAEI initiative (OAEI, 2009) is primarily a kind of competition amongst ontology matching algorithms. Since they mainly test automated matching routines, the results of their campaigns will not be used in the following chapters. However, for further development of the presented approach, they may be beneficial because multiple ontology matching is one of OAEI's forecasted targets yet undeveloped.

#### **2.4.3 Elicited ontologies**

The most recent form of ontology creation comes from the knowledge gained from tags in web 2.0 applications. They develop into elicited ontologies for classifications through interviews as in GI field (Mark, Smith, Tversky, 1999). The tagging by fiddling web users is taken as a source of information to elicit ontologies. Due to the different nature of interviews and web tags, several aspects, which are no longer stated before the interviews, have to be carefully considered. For instance, interviews are prepared through the selection of interviewed samples and questions to put to them. Web tags instead come freely without any previous selection from the interviewer. This research topic is gaining importance. We can see the rapid development from its inception (Gruber, 2007) (Mika, 2007), we rapidly obtained the first ontologies with simple formalizations (Knerr, 2008) (Kim, Scerri et al, 2008) followed by the federation of some of these to merge several aspects into one comprehensive

---

<sup>11</sup> <http://oaei.ontologymatching.org>

system (Kim, Passant et al, 2008). The latest development is the integration of folksonomies in the semantic web (Specia & Motta, 2007) (Angeletou et al, 2007) up to the most recent developments (Chen et al. 2010). The main features of elicited ontologies come from the social aspects (Gruber, 2008a) (Mika, 2007) and the contextual mining of tags. Since they are not developed by domain experts, a pivotal role is played by the construction or the recognition of a network of trusted actors to assess ontology usability (Bishr & Kuhn, 2007)<sup>12</sup>.

#### 2.4.4 Ontologic precision/expressivity

Ontologies can have many different levels of semantic expressivity and so many terms are used. A comprehensive classification, putting together definitions from both formal ontology literature and elicited social ontology literature, will be proposed in order to understand their relationship. From the formal ontology environment in (Obrst, 2006) the *Ontology spectrum* contains shared definitions and a classification for all the forms of knowledge representation as defined by the ONTOLOG community<sup>13</sup>. ONTOLOG comprises prominent experts in ontology research and applications. Another noticeable initiative of ONTOLOG is the *ontology framework* (Gruninger et al., 2008) which concerns the still ongoing definition of key dimensions to characterize ontologies. Obrst's explanation will be the backbone of the following paragraphs. Obrst underlines the difference between term and concept that will be used for finer distinctions. Terms are lexical forms indicating the meaning of each term. Concepts are expressed through the relationship between terms and play the role of a node or a link in a knowledge representation model. Another assumption Obrst made is that these kinds of knowledge representations are meant for formal ontologies, as described above.

Working in order of increasing magnitude of semantic expressiveness, the ontology spectrum classification is as follows:

- taxonomies
  - weak taxonomy
  - strong taxonomy
- thesauri
- conceptual model or weak ontology
- logical theory or strong ontology

Finer distinctions can be sourced (Smith & Welty, 2001), but the ONTOLOG definitions will be used due to the wide acceptance of this classification.

---

<sup>12</sup> OSM has developed a strong sense of community, the so called mapping parties, wiki pages and several mailing lists and chat rooms witnesses the existence of a strong community that is often stimulated by votes on proposal for new tags or on the election of the members of the foundation's board.

<sup>13</sup> ONTOLOG (a.k.a. "Ontolog Forum") is an open, international, virtual community of practice devoted to advancing the field of ontology, ontological engineering and semantic technology, and advocating their adoption into mainstream applications and international standards. The community portal of the community at <http://ontolog.cim3.net/>



At the first level a class taken from (Smith & Welty, 2001) can be inserted: the **catalogue**, which is simply a list without any form of hierarchy.

**Taxonomies** are classifications of terms and concepts in a tree like schema. They can be weak and strong according to the relationship they state between elements.

Weak taxonomies are hierarchies where terms and concepts are related through the relationship *subclassification*, while strong taxonomies make a distinction between concepts using the relationship *subclass of* which for terms are *narrower than* relationship. Taxonomies can be expressed formally using XML or RDF. Taxonomies are machine-readable.

**Thesauri** also known as controlled vocabularies refer to only terms and not concepts. In thesauri, you have four semantic relations between terms

Equivalence: synonymous terms.

Homographic: terms spelled the same.

Hierarchical: a term that is broader or narrower than another term (subsumption) is the semantic relationship we have already found in taxonomies.

Associative: related term (the relation is not further defined).

Thesauri can be expressed formally using XML schema and RDFS or database schema. Thesauri are machine-processible.

**Conceptual model or weak ontologies** define relationships between concepts similar to thesauri for terms that are well as represented by UML models and formally so by RDF schema (RDFS) (Brickley et al., 2004). Conceptual models are machine-processible.

**Logical theories or strong ontologies** can be either axiomatic or frame based according to whether the representation used is more entity focused or not. From this level the use of OWL for formalisation purposes starts. Logical theories are machine-interpretable. Finer subdivisions are outlined in the cited presentation but they are beyond the scope of this explanation that is more focused on making a clear distinction between terms used.

#### 2.4.5 Ontologies and folksonomies

Obrst has not classified elicited ontologies. What is the relationship between them and the formalized ones? According to (Mika, 2007) and (Gruber, 2007) folksonomies are ontologies cannot be classified in the (Smith & Welty, 2001) classification they refer to. Ontologies from folksonomies can be used to enrich formal ontologies that are reluctant to develop over time but also Mika underlines the fact that there are more semantics in folksonomies that can be extracted using network analysis that do not just add leaves to formally rigid structured trees. This is the approach of (Gruber, 2007b) and the ONTOLOG community (Gruninger, 2007) that give a pivotal role to formal ontologies. They sustain that social semantics can be augmented using a *snap to grid* approach aligning unstructured data to structured data (formalized ontologies). In (Mika, 2007) (Gruber, 2007a) (Gruber, 2007b) (Knerr, 2006) (Bishr & Kuhn, 2007) it is explicitly stated that folksonomies cannot be considered taxonomies (notwithstanding the origin of the name folks-

taxonomies) because of the absence of any explicit hierarchy in tags, therefore when unprocessed they should be considered open catalogues.

## 2.5 The environment of the proposed development: Linking Open Data

The first large-scale bootstrapping towards the semantic web is the web of linked data (Bizer et al., 2009)<sup>14</sup>. The Linking Open Data initiative is collecting, linking and publishing online data from every domain of human knowledge. To have an idea of the enormous amount of dataset collected until September 2010 we can refer to the following Figure 2-3

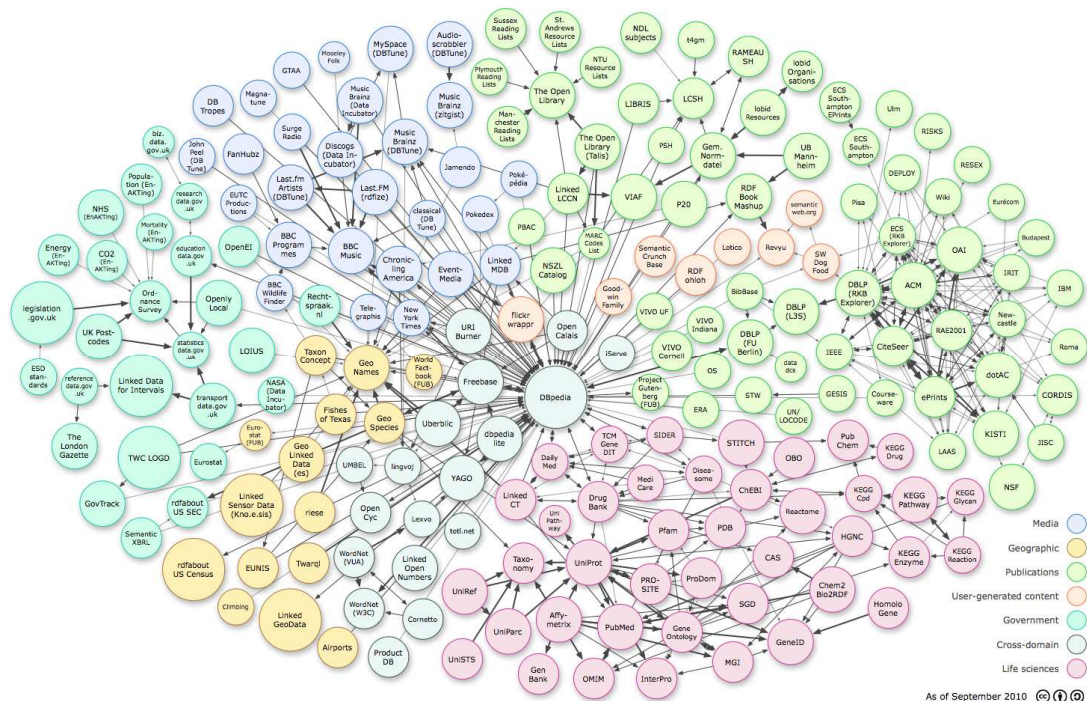


Figure: 2-3 The Linking Open Data cloud as of September 2010

In this picture, the 203 different datasets are coloured to allow the reader to distinguish between the different kinds of data origin. Data origin span from governmental data to user generated content, from media to publication. There is a specific category for Geographic data that collects geodata embracing all the diverse kinds of origins ranging from US census data (<http://www.rdfabout.com/demo/census/>) to linked sensor data ([http://wiki.knoesis.org/index.php/SSW\\_Datasets](http://wiki.knoesis.org/index.php/SSW_Datasets)) to geographic user generated content as in LinkedGeoData (<http://likedgeodata.org.>)

<sup>14</sup> A very interesting speech on it at [http://www.ted.com/talks/tim\\_berners\\_lee\\_on\\_the\\_next\\_web.html](http://www.ted.com/talks/tim_berners_lee_on_the_next_web.html) where Tim Berners-Lee the inventors of the web explains the value of Linked Data and uses OpenStreetMap as an example of valuable data online.

In this environment, a pivotal role is played by DBpedia (Auer et al., 2007). Moreover, semantic technologies can use the web of data in general without its results needing to be published on the web of data. Many tools are emerging to improve the value and usability of linked data (Bizer, Heath, Berners-Lee, 2009) and many of them may focus on the discovery of data either linked on the web or not. Linked data is the semantic web environment in which this thesis is developed. Compared to other alternatives the LOD allows for the integration of any form of knowledge collection without the need to have a local copy of it<sup>15</sup>. Several research projects worked on the integration of Geo resources and other forms of knowledge but they were limited and focused on the availability of local copies of data. With the rapid development of the LOD adding new open datasets at a daily rate, everything on the earth's surface will have the possibility to be mapped with several aspects of human knowledge as its nature. A brief introduction will be given to the Linking Open Data resources that are involved in the development of the thesis.

### 2.5.1 DBpedia

DBpedia is the initiative to extract information from Wikipedia and to convert it in a structured semantic manner on the web. It is a cornerstone for the web of data (Bizer, Lehmann, et al., 2009). It has been fostered mainly in the German academic environment with the leading universities of Leipzig and Berlin. DBpedia has been developed using OWL DL. Therefore, equivalence is stated at instance level and not at class level.

### 2.5.2 Geodata in DBpedia

Looking at the DBpedia ontology *Figure: 2-4 Places in DBpedia ontology*, on the following page, have been extracted. It shows the class hierarchy used to represent elements in the earth's surface. Due to its origin, DBpedia lacks a correct classification hierarchy, as all folksonomies need to be integrated with upper level or well-formalized ontologies. In the following Figure 2-4 you can see how the structure lacks a unifying conceptual model of reality as required for any collection of geographical information as in (Peuquet, 1998) and in (OGC,2009). Moreover, it is missing common sense: islands, continents and countries are classified as populated places. The classification in DBpedia needs to be improved to create considerably more abstraction levels and should at least be compared to common linguistic resources like WordNet as the second abstraction level, amongst the nine in (OGC, 2009), is the conceptual world where distinctions are based on natural language.

---

<sup>15</sup> Recently this possibility has been weakened by the maintainers of the main access point to query the LOD data the SPARQL endpoint at (<http://lod.openlinksw.com/sparql>). They reduced the time for having answers. Consequently, the majority of queries of the present work, that lasted hours, are no more possible.



several contributions and grew rapidly and efficiently (Miller, 2007). In 2007 version 3.0 was issued after several intermediate versions.

*WordNet, an electronic lexical database, is considered to be the most important resource available to researchers in computational linguistics, text analysis, and many related areas. Its design is inspired by current psycholinguistic and computational theories of human lexical memory. English nouns, verbs, adjectives, and adverbs are organized into synonym sets, each representing one underlying lexicalized concept. Different relations link the synonym sets.* (Fellbaum,1998).

The project is very wide-ranging and tries to explain the way people use language and cognitive relationships. Returning to the objectives of the present thesis and to explain what is most often used, a brief description of the linguistic formalization and terminology used in WordNet follows.

### 2.6.1 WordNet data model

Words in WordNet are mainly classified in *synsets* that include all the words that can be intended as synonymous, having in certain contexts the same word sense. It means that a word can be part of more than one synset. There are synsets for nouns, verbs, adjectives and adverbs. Other kinds of relationships between words and synsets are as follows:

*WordNet defines seventeen relationships, of which ten between synsets (hyponymy, entailment, similarity, member meronymy, substance meronymy, part meronymy, classification, cause, verb grouping, attribute) and five between word senses (derivational relatedness, antonymy, see also, participle, pertains to). The remaining relationships are "gloss" (between a synset and a sentence), and "frame" (between a synset and a verb construction pattern).* from (Assem , Gangemi, Schreiber ,2006)

There are several relationships between synsets according to their nature of being noun, verb, or adjective. Focusing on nouns, the relationship between synset that pertain to nouns are as follows. From WordNet online glossary<sup>16</sup>

*Two kinds of relations are represented by pointers: lexical and semantic. Lexical relations hold between semantically related word forms; semantic relations hold between word meanings. These relations include (but are not limited to) hypernymy/hyponymy (superordinate/subordinate), antonymy, entailment, and meronymy/holonymy.*

Nouns and verbs are organized into hierarchies based on the hypernymy/hyponymy relation between synsets.

In detail:

**antonym** thus the opposites e.g. the following couples: male female, long short, up and down.

**entailment**, or logical implication if B is true it means that also A is true

**meronymy** is the semantic relation between an object A and the object B when A is a part of B e.g. a finger is part of a hand

---

<sup>16</sup> (<http://wordnet.princeton.edu/wordnet/man/wngloss.7WN.html>)

**holonymy** is the semantic relation between an object and a part of it e.g. hand and fingers, it is the opposite of meronymy so similarly to the previous example B is meronymy of A so a finger is a meronymy of a hand while a hand is a holonymy of fingers

**hypernymy** expresses the subclass of the relation between two entities e.g. scarlet is a type of red then the semantic relationship between scarlet and red is that scarlet is a hypernymy of red. In other words scarlet is a subclass of red.

**hyponymy** is the opposite of hypernymy e.g. red is a hyponymy of scarlet thus we can also say that red is a superclass of scarlet.

Returning to our thesis, if someone is querying the OSM database, apart from synonymy one could think about forming extended queries in more general terms including objects whose queried object is a part (then meronymy) or objects whose queried term is a subclass (then hypernymy). In the development of the present work, the focus will only be on synonymy.

### 2.6.2 Previous applications of WordNet in GI

Some projects have already used WordNet in the GI environment to enhance the query mechanism. Sometimes it has been used to infer relationships between places, in (Buscaldi et al. 2006) it is used to expand queries using the meronymy/holonymy relation that in the GI environment is also called partonomy.

Sometimes it has been used to integrate different GI databases like a semantic component in a considerably more complex procedure (Buccella et al, 2009) (Buccella et al, 2010). The cited work is quite similar to the present thesis, but it does not rely on the Linking Open Data, it is a standalone application not connected to the web of data. The latest work from Buccella, requires development to broaden the linguistic relations between objects; moreover, it often involves programming.

### 2.6.3 WordNet in the semantic web

Due to the importance of WordNet, several research projects focused on its translation in RDF to embrace it in the semantic web. The most mature of them developed in 2006 is available on the semantic web, through a W3C initiative (Assem, Gangemi, Schreiber, 2006) that translated the 2.0 version in RDF/OWL. Furthermore, other projects developed the 3.0 WordNet version<sup>17</sup> though it is still not available on the semantic web<sup>18</sup>. The RDF/OWL representation of WordNet is online and can be queried through two SPARQL endpoints. The first one is the RKBEXPLORER that contains WordNet RDF/OWL only accessible at <http://wordnet.rkbexplorer.com/sparql/>. WordNet RDF/OWL can also be queried through the LOD SPARQL endpoint <http://lod.openlinksw.com/sparql> so it can be queried simultaneously with LinkedGeoData, which is the semantic translation of OpenStreetMap.

---

<sup>17</sup> <http://semanticweb.cs.vu.nl/lod/wn30/>

<sup>18</sup> The publication already started but it did not worked properly when tested.

The WordNet knowledge base is structured as follows:

- Synset
  - AdjectiveSynset
    - AdjectiveSatelliteSynset
  - AdverbSynset
  - NounSynset
  - VerbSynset
- WordSense
  - AdjectiveWordSense
    - AdjectiveSatelliteWordSense
  - AdverbWordSense
  - NounWordSense
  - VerbWordSense
- Word
  - Collocation

The class hierarchy and the relationship between the main concepts of the WordNet schema from (Assem, Gangemi, Schreiber, 2006) in the following Figure 2-5:

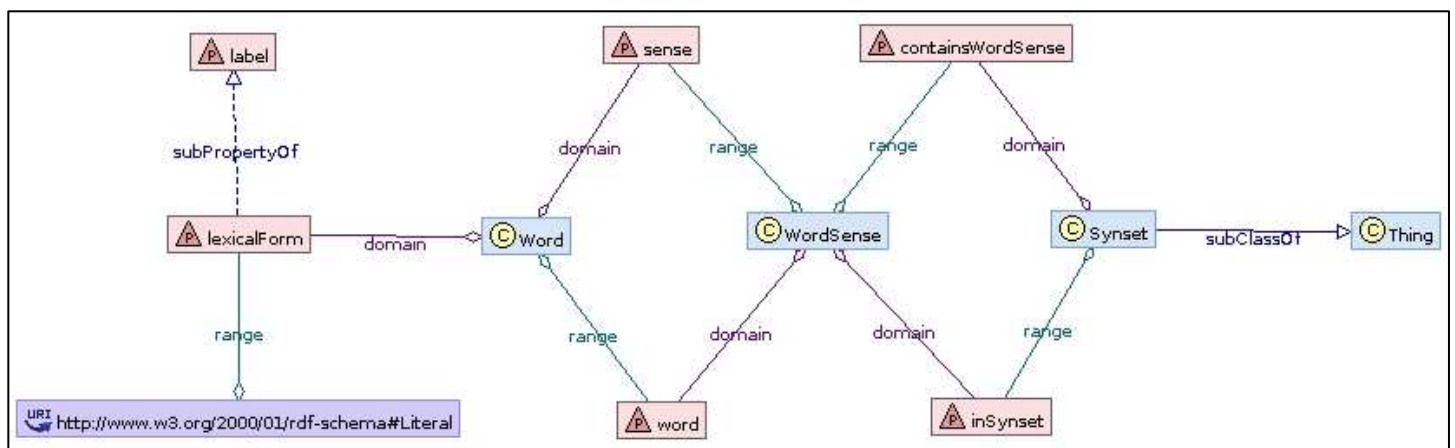


Figure: 2-5 WordNet RDF/OWL class hierarchy. Using RDF Gravity

The instances are called lexical forms. A lexical form is a word whose word sense is included in a synset. This structure will often be used in the queries over the LOD.

## 2.7 Folksonomies and the semantic web

OSM is a GI folksonomy and therefore can be seen through the lens of the rich spectrum of research on folksonomies in the SW. Several terms have been used to define social semantic collective intelligence or collected intelligence. Gruber (2007) thinks that the killer application for the SW will

come from applications that mix user generated content with formal semantics. To reuse, model and gain more semantics than explicitly stated in folksonomies, several attempts has been proposed and a plethora of projects are ongoing (Specia & Motta, 2007). A comparative study (Kim, Scerri et al., 2008) and an alignment proposal amongst some of them in the SW (Kim, Passant et al. 2008) look promising for further development of the present work. Since the existence of some kind of metadata regarding time and editors are prerequisites for all development from folksonomies to ontologies, both OpenStreetMap and LinkedGeoData is equipped with some kind of metadata such as the user and the changeset so the social and the temporal dimensions are linked with the positional<sup>19</sup>.

The alignment proposal tries to expand the value of tags interlinking various SW resources like the following:

- Ontologies developed for tagging.
- SIOC<sup>20</sup>, for the management of social network information.
- FOAF<sup>21</sup> for personal information.
- SKOS<sup>22</sup> to organize collection of tags.
- DC<sup>23</sup> for management of metadata.

Surprisingly, there is not an ontology extracted from folksonomies that relies on WordNet. Integrating it by means of the merged approach suggested in (Kim, Passant et al. 2008) would be beneficial both for the semantics embedded in tags and for possible extensions of WordNet. In (Chen et al., 2010), pre-processing of data to extract basic level concepts proved to work well, but only in a standalone environment. A similar approach could be developed in the LOD with the possibility of gaining the advantage of well structured cognitive and linguistic resources like DOLCE and WordNet. In a broader perspective, WordNet itself could take advantage of folksonomies since they are rich in jargon and new terminology (Angeletou et al., 2007). Creating a system to harvest new words and concepts from tags can facilitate the updating process and the expansion of WordNet. For instance, we will see that in WordNet many terms that are used to describe transportation networks are intended as synonyms while in OSM jargon they are differentiated. In this case, the different assumptions would weaken the chosen approach of the present thesis.

---

<sup>19</sup> Not spatial because no topology is embedded in OSM and LGD

<sup>20</sup> [Sioc-project.org](http://sioc-project.org/)/ The SIOC initiative (Semantically-Interlinked Online Communities) aims to enable the integration of online community information.

<sup>21</sup> <http://www.foaf-project.org/> The *Friend of a Friend* (FOAF) project is creating a Web of machine-readable pages describing people, the links between them and the things they create and do

<sup>22</sup> <http://www.w3.org/2004/02/skos/> SKOS is an area of work developing specifications and standards to support the use of knowledge organization systems (KOS) such as thesauri, classification schemes, subject heading lists and taxonomies within the framework of the Semantic Web

<sup>23</sup> Dublin Core metadata initiative <http://dublincore.org/> in the LOD are hosted by PURL project



## **2.8 Maturity of ontology development**

Although recent developments in ontology engineering show advantages of reusable and flexible design practices, such as Ontology Design Patterns (Gangemi, 2005), the ontologies developed for OSM are poorly structured, and have been developed following a waterfall procedure. To evaluate the maturity of the ongoing ontology creation efforts for OpenStreetMap they will be compared with ontological engineering according to the waterfall procedure designed by Uschold and Grüninger (1996). The ontology design process follows the steps:

- 1 Identify Purpose and Scope*
- 2 Building the Ontology*
  - 2.1 ontology capture,*
  - 2.2 ontology coding,*
  - 2.3 integrating existing ontologies*
- 3 Evaluation*
- 4 Documentation*
- 5 Guidelines for each phase.*

## 3 Chapter three: Geoweb 2.0 and OpenStreetMap

### Introduction

In the following chapter the selected VGI application, OpenStreetMap will be illustrated. After a short comparison with the other VGI initiatives to support the choice of OSM in section 3.1; there will be a short explanation on the data structure and format in sections 3.3 and 3.4, the working of the community and the database structure in section 3.5; finally the tools that have been developed to extract collected geodata in section 3.6.

### 3.1 OpenStreetMap and VGI

OpenStreetMap (OSM) is the most successful VGI initiative, describing itself as the Wikipedia for maps. Started in 2004 from an idea by Steve Coast<sup>24</sup> with a rapidly growing community of almost 290,000 users registered as of July 2010<sup>25</sup> it has a growing global coverage and an enormous amount of vector data that has been created and updated. The database has reached almost 200 GB. On the following page in *Figure: 3-1 The growth of OpenStreetMap - users and points* from 2005 to the end of July 2010.

Data is collected in many ways. Initially, location data was gathered mainly through GPS, afterwards the project developed adding the possibility to gather information from free satellite imagery, or importing from freely available vector datasets. The map can be viewed online at <http://www.openstreetmap.org>.

Data is downloadable free of charge and this free re-usability of data drove the development of the project:

*The project was started because most maps which are thought of as free actually have legal or technical restrictions on their use, preventing people from using them in creative, productive, or unexpected ways.*<sup>26</sup>

Surveying is a considerably ancient activity; therefore, it is also possible without a GPS. In Cuba where GPS usage is not allowed they use classical surveying techniques. Furthermore, the enthusiastic community is developing more and more tools to improve the experience of mapmaking.

---

<sup>24</sup> In 2009 voted like the second most influential Geospatial Person by readers of an online geo magazine [http://www.directionsmag.com/article.php?article\\_id=3225](http://www.directionsmag.com/article.php?article_id=3225)

<sup>25</sup> <http://wiki.openstreetmap.org/wiki/Stats>

<sup>26</sup> [http://wiki.openstreetmap.org/wiki/Beginners%27\\_Guide](http://wiki.openstreetmap.org/wiki/Beginners%27_Guide)

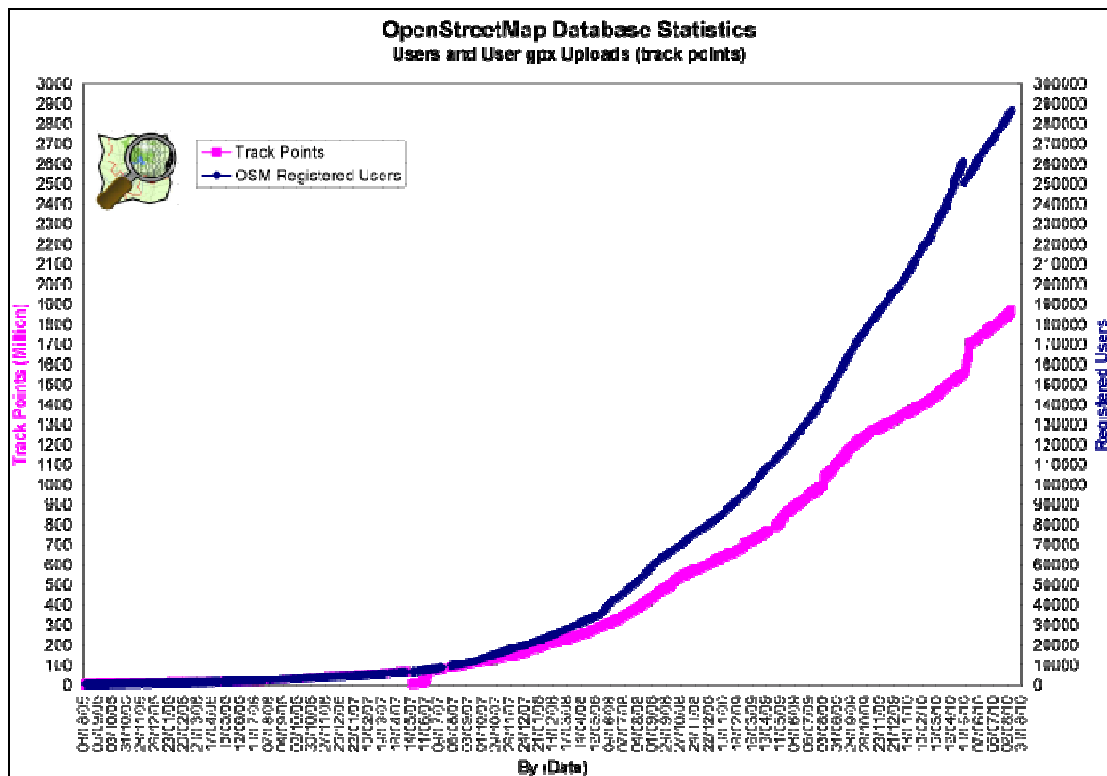


Figure: 3-1 The growth of OpenStreetMap - users and points

Peer production of Geographic content in OSM differs from the others because:

- Flickr<sup>27</sup> is more oriented towards tagging pictures, it is a mashup using Google Maps.
- Wikimapia is more oriented towards tagging entities and is a mashup using Google Maps.
- Way Faring<sup>28</sup> supports the creation of maps adding POI through a mashup using Google Maps.
- Google Mapmaker<sup>29</sup> is also oriented towards creating and tagging entities.
- OpenStreetMap<sup>30</sup> is oriented towards creating and tagging entities.

The main difference between OSM and Google Mapmaker is the restriction that Google poses on the use of its data, the restricted number of users and the restricted geographic coverage. In OSM, any contribution from users is welcome and everyone can map whatever he/she likes. From the initial focus on streets, at present anything can be mapped depending on users' willingness in a far more detailed way compared to commercial datasets available online.

<sup>27</sup> <http://www.flickr.com>

<sup>28</sup> <http://www.wayfaring.com>

<sup>29</sup> <http://www.google.com/mapmaker>

<sup>30</sup> <http://www.openstreetmap.org/>

Following figures 3-2, 3-3, and 3-4 for some comparison between Google Maps (left side) and OSM<sup>31</sup> (right side).

In the first case, even if the quality is comparable, OSM provides its own rendering of small channels and pedestrian paths, as to more detailed buildings and road networks. The differences are noticeable in the recreation areas where commercial companies have little interest in detailing.

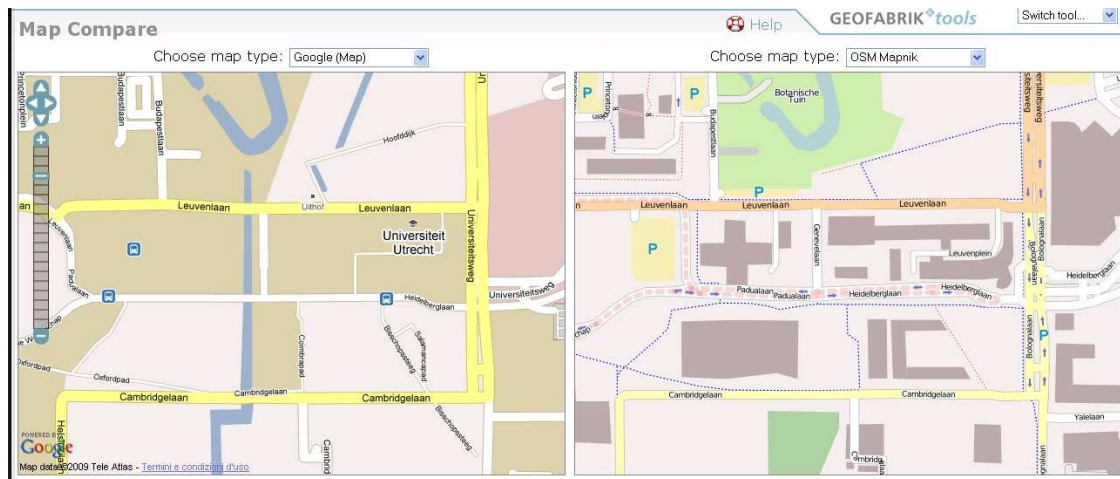


Figure: 3-2 Google Maps and OSM comparison - De Uithof

Figure 3-3 shows a capture focusing on Amsterdam Zoo<sup>32</sup>.

In Google maps (left side), you have a green indefinite area while in OpenStreetMap (right side) you have a detailed map where the park is rendered with paths and facilities and the location of the different animal enclosures.

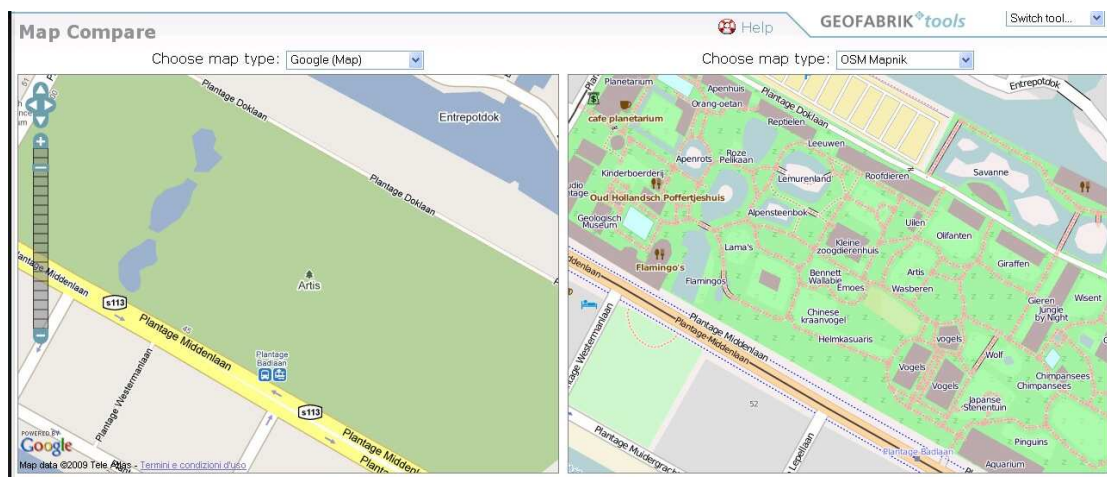


Figure: 3-3 Google Maps and OSM comparison - Amsterdam Zoo.

One would think commercial projects would have a better coverage than

<sup>31</sup><http://tools.geofabrik.de/mc/?mt0=mapnik&mt1=googlemap&lon=5.17199&lat=52.08521&zoom=16>

<sup>32</sup><http://tools.geofabrik.de/mc/?mt0=mapnik&mt1=googlemap&lon=4.91638&lat=52.36609&zoom=17>

VGIs, but if we examine the coverage of the island of Cyprus in figure 3-4 we can see how that assumption is wrong<sup>33</sup>.

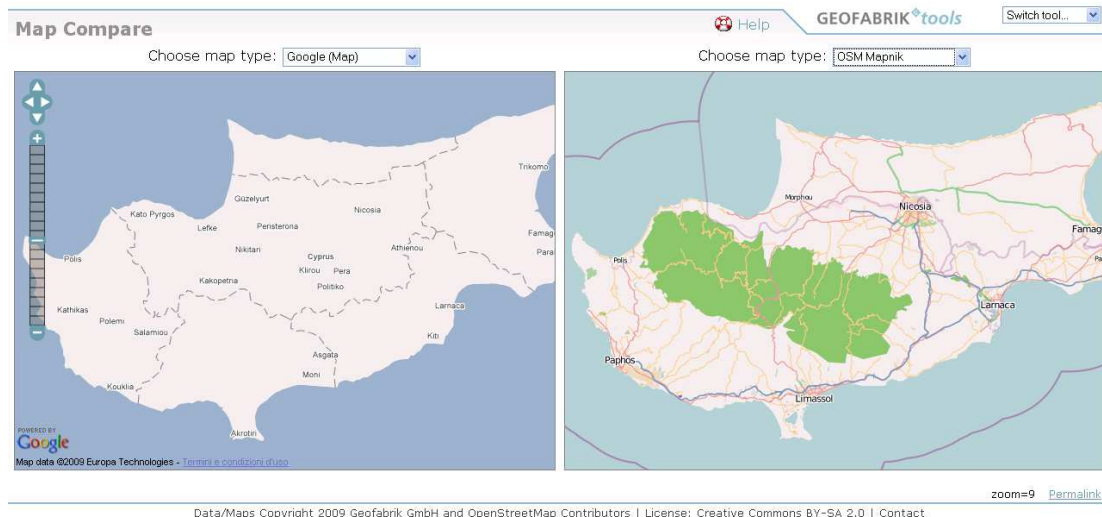


Figure: 3-4 Google Maps and OSM comparison – Cyprus

Finally, due to the freedom of surveying that OSM allows, the database contains considerably more information than the render engines are able to display. In the following paragraphs, a full description of all the blocks of the OSM project will be avoided. This task is beyond the objective of the thesis and some good explanations are in the cited literature. The focus will be only on the parts influencing the research object.

### 3.2 Data in OpenStreetMap

Data is positional data and tags. Spatial entities (data primitives) are nodes ways/areas and relations<sup>34</sup>. There is not a distinction between real world objects and features, as in GML, therefore a way can be either a street or a border. Ways are series of nodes; areas are closed ways while relations are a collection of nodes and ways that share some characteristic. Relations are also used for multipolygon definition. Tags are defined by a series of key=value where a specific key can be defined only once for an entity while several keys can be associated to a single entity. A set of suggested pairs key/values is in the Map features<sup>35</sup> webpage of project wiki.

OSM was structured to make people contribute as easy as possible; therefore, they refrained from using almost all GIS standards<sup>36</sup> while they fully embraced IT web standards.

The datum is the WGS-84 only due to the massive use of GPSs.

<sup>33</sup> <http://tools.geofabrik.de/mc/?mt0=mapnik&mt1=googlemap&lon=33.2454&lat=35.08014&zoom=9>

<sup>34</sup> [http://wiki.openstreetmap.org/wiki/Data\\_Primitives](http://wiki.openstreetmap.org/wiki/Data_Primitives)

<sup>35</sup> [http://wiki.openstreetmap.org/wiki/Map\\_Features](http://wiki.openstreetmap.org/wiki/Map_Features)

<sup>36</sup> [http://wiki.openstreetmap.org/wiki/FAQ#Questions\\_from\\_GIS\\_people](http://wiki.openstreetmap.org/wiki/FAQ#Questions_from_GIS_people)

### 3.3 The OSM structure

The project is composed mainly of the input storage and output parts as shown in figure 3-5. In the input component, the construction of the community is stressed. Mapping parties (above all in Figure 3-5), are meetings where people go around to collect geoinformation of a selected area. Later agreements on tags to use are created through discussions on the wiki and chats. The more technical part of input is left to software that can combine different sources of data to help the making of maps. Users can upload GPS traces to guide the editing; users can also import vector data or trace maps from background pictures often coming from WMSes (green box in Figure 3-5). All vector data imported or raster data traced have to be compliant with the Creative Commons Attribution-Share Alike 2.0 Generic licence<sup>37</sup> shortly CC BY SA. All data imported have to pass through the project API that store edits in the database. The output of data (blue box in figure 3-5) can be through dumps of the database either through the API or through the OSMOSIS application. However, the main purpose of the output part is the creation of the tiles that are rendered in the project's website. The DBMS moved on 19 April 2009 from MySQL to PostgreSQL server with the PostGIS extension used for rendering purposes. In Figure 3-5, over a picture of OSM components taken from the wiki<sup>38</sup> the main parts of the project have been sketched using colored boxes.

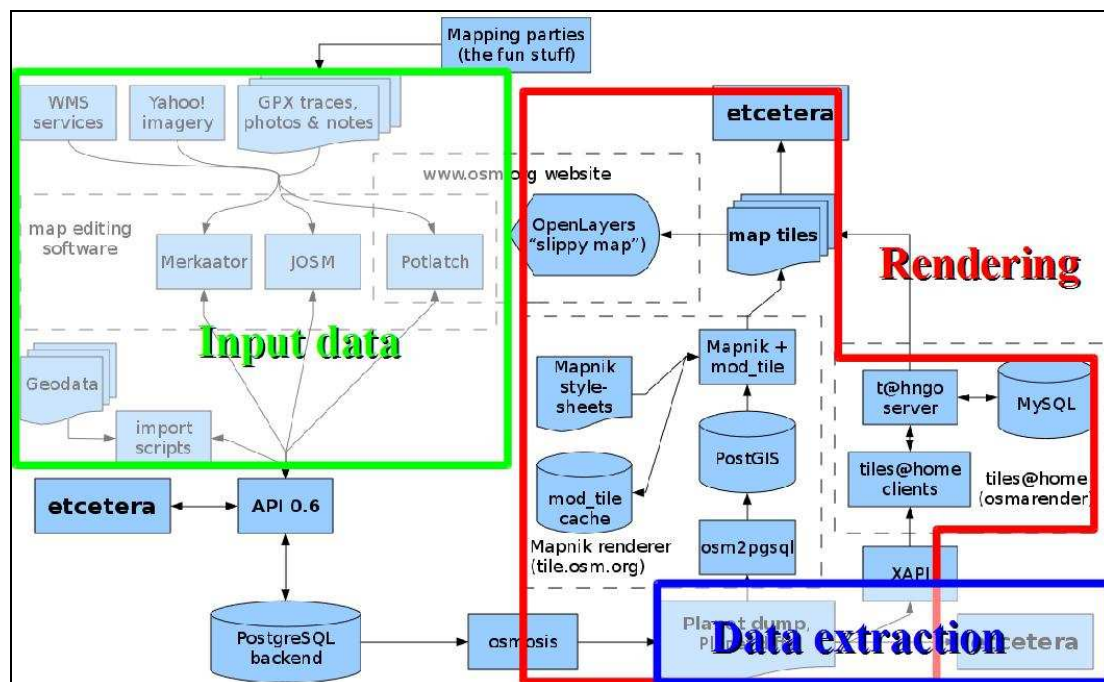


Figure: 3-5 OSM components and parts

<sup>37</sup> <http://creativecommons.org/licenses/by-sa/2.0/>

<sup>38</sup> [http://wiki.openstreetmap.org/wiki/Component\\_overview](http://wiki.openstreetmap.org/wiki/Component_overview)

### 3.4 Data format

The data format is XML formatted and represents all data primitives and notions about the changesets and users that have contributed to the element.<sup>39</sup> It is released using the .osm format. The differences between the terms 'primitives/elements' that may look synonymous must be emphasized.<sup>40</sup> Elements are the rendered elements in OSM maps, while primitives pertain to the data representation in the database<sup>41</sup>.

#### Nodes

Node is the main primitive; the other primitives are a collection of nodes. Below is an example of a node with tags describing the position, the user that created it or updated it last, the time of last modification, a single key/value pair, the visibility tag and the software used for editing (the automated insertion of the last two tags is no longer supported):

```
<node id="25496583" lat="51.5173639" lon="-0.140043" user="80n"
visible="true" timestamp="2007-01-28T11:40:26+00:00">
  <tag k="created_by" v="JOSM"/>
</node>
```

Listing: 3-1 OSM primitives node example

#### Ways

Ways are a collection of nodes.

Following is an example of a way in the OSM database; numbered lines are added for explanation later:

```
1. <way id="5090250" visible="true" timestamp="2008-05-
03T12:16:45+01:00" user="Andy Allan">
2.     <nd ref="822403"/>
3.     <nd ref="21533912"/>
4.     <nd ref="821601"/>
5.     <nd ref="21533910"/>
6.     <nd ref="135791608"/>
7.     <nd ref="823771"/>
8.     <tag k="oneway" v="yes"/>
9.     <tag k="highway" v="unclassified"/>
10.    <tag k="created_by" v="Potlatch 0.8c"/>
11.    <tag k="name" v="Clipstone Street"/>
12. </way>
```

Listing: 3-2 OSM primitives way example

Here another tag explains which software has been used to edit the element (potlatch is the online editor). Areas are modelled similarly, they are not

<sup>39</sup> <http://wiki.openstreetmap.org/wiki/.osm>

<sup>40</sup> <http://wiki.openstreetmap.org/wiki/Elements>

<sup>41</sup> The following XML examples are taken from [http://wiki.openstreetmap.org/wiki/Data\\_Primitives](http://wiki.openstreetmap.org/wiki/Data_Primitives)

another primitive, they are simply ways with coinciding starting and ending nodes. In the example above, we might have another line identical to line 2 after line 7.

## Relations

Relations are used to model relationships between entities or to express abstract objects such as boundaries. Following is an example of relation:

```
<relation id="12" timestamp="2008-12-21T19:31:43Z"
user="kevjs1982" uid="84075">
  <member type="way" ref="2878061" role="outer"/>
  <member type="way" ref="8125153" role="inner"/>
  <member type="way" ref="8125154" role="inner"/>
  <member type="way" ref="3811966" role=""/>
  <tag k="created_by" v="Potlatch 0.10f" />
  <tag k="type" v="multipolygon" />
</relation>
```

Listing: 3-3 OSM primitives relation example

In this case, the relation expresses holes in an area. As you may have noticed, even the simplest element holds contextual information on who has tagged and when the tagging happened in the dataset. It provides a kind of metainformation on the tagging process that together with the strong sense of community makes OSM data more reliable for data analysis and governance of the tagging process as suggested by (Bishr & Kuhn, 2007).

## 3.5 Database

The database<sup>42</sup> stores everything related to the project, from geometric to attribute values to user preferences that are only displayed in the wiki. The database stores all the history of the elements. Two tables for each data primitive are taken: the current is the front end for reading, writing, rendering and extracting data while the master table contains all previous versions and is not downloadable. Given the wiki style of the project developers pages are freely accessible and the latest versions of database structure are accessible<sup>43</sup>. Below a list of tables is provided, focusing only on current geometric attributes and tags. The first tuples in bold is the name of the table, while the following tuples refers to columns of the table. The rows “k” and “v” refer to Key and Value, the tagging couples. In the database there is also the master table (the old version) maintained to hold the history or to revert to a previous status of the database to avoid vandalism damage or to reject the import of data that does not comply with the license. The master table is

<sup>42</sup> [http://wiki.openstreetmap.org/wiki/Database\\_schema](http://wiki.openstreetmap.org/wiki/Database_schema)

<sup>43</sup> <http://git.openstreetmap.org/?p=rails.git;a=tree;f=db/migrate>



identical to the one depicted below with the only difference being that the name of the table is without the prefix “current\_” in the name.

## NODES

The main primitive, you can recognize the elements found in *Listing: 3-1 OSM primitives node example*.

<b>current_node_tags</b>
id
k
v

Table: 3.1 OSM tags of nodes as currently rendered.

<b>current_nodes</b>
id
latitude
longitude
changeset_id
visible
"timestamp"
tile
version

Table: 3.2 Metadata and location of current nodes in OSM

## WAYS

Ways have similar tables with the first outlining the sequence of nodes that constitutes the ways, in the following you can recognize the elements found in *Listing: 3-2 OSM primitives way example*.

<b>current_way_nodes</b>
id
node_id
sequence_id

Table: 3.3 OSM ways through nodes sequence

<b>current_way_tags</b>
id
k
v

Table: 3.4 OSM tags for ways

<b>current_ways</b>
id
changeset_id
"timestamp"
visible
version

**Table: 3.5 OSM ways metadata**

## RELATIONS

Similarly to ways, relations rely on a coordinating table identifying members of each relation stating their type since relations can involve both nodes and ways. Moreover, you can read the purpose in the relation for every element (member\_role). In the following, you can recognize the elements found in *Listing: 3-3 OSM primitives relation example*.

<b>current_relation_members</b>
id
member_type
member_id
member_role
sequence_id

**Table: 3.6 OSM relation members**

<b>current_relation_tags</b>
id
k
v

**Table: 3.7 OSM relation tags**

<b>current_relations</b>
id
changeset_id
"timestamp"
visible
Version

**Table: 3.8 OSM relations metadata**

### 3.6 Tools to extract data

Several options to extract data are available; to extract parts of the dataset we have two main tools:

From the map view through the API<sup>44</sup>

Minute changes pass through OSMOSIS<sup>45</sup> that is a community developed Java tool.

The other extraction facilities rely on those two.

To extract the full database the planet.osm<sup>46</sup> file is available.

Extracts of the planet file for example at national level are available online.

All the above mentioned data and settings pertain to the XML based environment where semantic is excluded.

---

<sup>44</sup> [http://wiki.openstreetmap.org/wiki/Downloading\\_data](http://wiki.openstreetmap.org/wiki/Downloading_data)

<sup>45</sup> <http://wiki.openstreetmap.org/wiki/Osmosis>

<sup>46</sup> <http://wiki.openstreetmap.org/wiki/Planet.osm>

## 4 Chapter four: OpenStreetMap and the semantic web

### Introduction

This chapter is entirely devoted to the analysis of initiatives attempting to implement new semantic technologies either inside or outside the general framework of OpenStreetMap (sections 4.2, 4.3, 4.5, 4.6). Amongst all the initiatives, LinkedGeoData will be analyzed more in details (section 4.4). LinkedGeoData was used for the development of the thesis as the semantic translation of OpenStreetMap. As a framework for evaluation, an initial section (4.1) is devoted to the description of the features a semantic geographic resource might have. At the end of the chapter, a comparison between the semantic initiatives is shown in section (4.7).

### 4.1 Geographic Information ontologies

Several projects have used semantic technologies in the GI field as mentioned above. To establish a framework to evaluate the ongoing geo-ontology efforts developed in the OSM environment, two approaches have been considered. The first has been developed focusing on the features a formalized domain GI ontology might have (Tomai & Kavouras, 2004). Most of the listed characteristics can be associated with any field of knowledge. Focusing on geoinformation, it stresses the points relating to spatial location and spatial relations that are domain specific; but some characteristics like semantic relations are shared with all fields of knowledge, for those concepts a meeting stage can be found in upper level ontologies.

In table 4.1 on the following page, the main elements and the description a geographic information ontology might have according to (Tomai & Kavouras, 2004).

A more specialized approach was developed in the SPIRIT research project<sup>47</sup> whose objective was *the design and implementation of a search engine to find documents and datasets on the web relating to places or regions referred to in a query*. In (Abdelmoty, et al, 2005) a geo ontology for a semantic search engine was developed. OWL proved itself more flexible than GML in fulfilling the designing of a geo-ontology for SPIRIT.

The elements of the geo-ontology for the SPIRIT search engine are:

- Actual and alternative place names
- Geographical containment hierarchies
- Place types
- Footprints (multiple spatial representations of geographic places)

In the list above we can find all the characteristics listed in table 4.1.

---

<sup>47</sup> <http://www.geo-spirit.org/>

The possibility to identify features with alternative names is only more explicitly expressed.

GI ontology elements	Explanations and contextualization in GI
Concepts	Concepts involved in the scope of the ontology or coming from existing domain ontologies as outlined for domain ontologies in 2.4.1.
Lexicon or vocabulary	Definition of concepts and relations in natural language plus metadata on how the ontology might be used.
Relations as semantic relations	Relations between concepts at semantic lexical level: Synonymy, hyperonym /hyponymy, meronym/holonym they pertain to upper level ontologies.
Relations as semantic property	Properties for concepts and values they can take according to the task of the ontology: spatiality, temporality, nature, material/cover, purpose, activity. Furthermore, the roles means and results of the specific activity might be specified.
Relations among relations	Taxonomy of relations between semantic relations and semantic properties.
Relations as axioms	Constraints on concepts and properties to assess consistency and completeness of the ontology: metrology, location, topology.

**Table: 4.1 Geographic information ontology elements**

Relying on the web of data, the integration of geoinformation with primarily lexical information can be achieved through a mapping with WordNet RDF/OWL since, as seen in 2.6.1, in WordNet semantic relations are already implemented. GI ontologies, with all the attributes, according to (Tomai & Kavouras, 2004) in the context of the ontology framework as stated in ONTOLOG community (Gruninger et al., 2008) and previously described in 2.4.4 are placed amongst the higher levels of expressivity.

## 4.2 Semantic efforts and OpenStreetMap

Several attempts have been carried out to introduce new semantic tools related to OSM. Attempts followed two different approaches, trying either to improve the ability of “users” as creators of data or to support the “user” as the data consumer. Firstly, attempts to improve the project internally have been analyzed; tasks discussed in the official mailing list and reported in the wiki are intended to improve the wiki and the mapmaking process. They are:

- The tag central proposal<sup>48</sup>; to give the tags a structure and meaning.

<sup>48</sup> <http://www.frankieandshadow.com/sotm10/tagcentral.pdf>

The proposed evolution depends on the creation of a more complex database. It provides semantic without semantic technologies and without semantic web. For the above cited limits, it will be ignored in further development of the thesis. It is another effort to explain things that in a linked open environment can be developed avoiding the duplication of semantic efforts.

- The attempt to implement a semantic mediawiki<sup>49</sup>. It encompasses the creation of a machine-readable map feature list<sup>50</sup>.
- At a draft stage there is an attempt to create an OWL description of SVG maps created for osmarender, <sup>51</sup> one of the most common OSM rendering machines<sup>52</sup>.

The second kind of semantic project relies on external projects trying to use the data collected by OpenStreetMap. They are:

- The massive translation of the database in RDF and the linking of the resulting knowledge base with DBpedia<sup>53</sup> have been carried out in the LinkedGeoData (following LGD) initiative<sup>54</sup> (Auer, Lehman, Hellman, 2009).
- The implementation of a XSLT stylesheet to convert the .osm files directly in RDF triples<sup>55</sup> has been carried out by a user (Simon Reinhardt).

In the following, the above cited initiatives have been analysed and their methodology has been evaluated as interesting case studies or resources to use for the present thesis or for further developments.

### 4.3 The Semantic MediaWiki – Machine readable map feature list effort

The Semantic Media Wiki implementation is the only semantic task listed in the “Things to do” page in OSM wiki<sup>56</sup>

Nowadays the tags listed in the map feature page have also some wiki pages with additional information like a short description, the group the tag belongs to, which primitives can be tagged this way, useful combinations with other tags, tags implied by the use of the considered one, figures and usage statistics. In transforming the wiki semantically, the community tends to automate the creation of the map feature list from all the tag description pages<sup>57</sup>. One of the listed advantages is that “*apps can use the wiki as data source, caused by simple data export*” in RDF format. The taxonomy can thus

---

<sup>49</sup> [http://wiki.openstreetmap.org/wiki/Semantic\\_MediaWiki](http://wiki.openstreetmap.org/wiki/Semantic_MediaWiki)

<sup>50</sup> [http://wiki.openstreetmap.org/wiki/Machine-readable\\_Map\\_Feature\\_list](http://wiki.openstreetmap.org/wiki/Machine-readable_Map_Feature_list)

<sup>51</sup> <http://wiki.openstreetmap.org/wiki/User:Esscue/OsmarenderOWL>

<sup>52</sup> The default style sheet for rendering is Maplink

<sup>53</sup> <http://dbpedia.org/About>

<sup>54</sup> <http://linkedgeodata.org/About>

<sup>55</sup> <http://osm.bloody-byte.net/documents/index.html>

<sup>56</sup> [http://wiki.openstreetmap.org/wiki/He:Things\\_To\\_Do#Machine\\_Readable\\_Feature\\_List](http://wiki.openstreetmap.org/wiki/He:Things_To_Do#Machine_Readable_Feature_List)

<sup>57</sup> [http://wiki.openstreetmap.org/wiki/Semantic\\_MediaWiki](http://wiki.openstreetmap.org/wiki/Semantic_MediaWiki)

be easily exported or linked with other applications. In another page related to this development “*search in tag definitions, through keys, possible values and (localized) descriptions of tags (Example: user wants to tag a children’s<sup>58</sup> playground. He is not sure whether the appropriate tag is amenity=playground, leisure=playground leisure=childrens\_playground. He searches for 'children’s playground' and gets leisure=playground as a result)*” is stated.<sup>59</sup> Unfortunately, nothing to perform such a task has been found in the documentation. Because of the desire to work only on accepted map features, this effort is leaving out all the tags that are not listed and documented. It restricts the potential of the semantic enhancement. Ontology has not continued to develop; latest updates in the wiki are dated August 2009.

## 4.4 LinkedGeoData

### 4.4.1 Overview

LinkedGeoData is the only working attempt to semantically translate OSM<sup>60</sup> fully.

Developers, starting from their IT background tried to optimize the import of OSM data in the semantic web assuming a couple of departures from an uncritical translation of OSM tags in RDF.

The project has been developed as follows:

- Firstly, developers optimized tools to extract data from OSM database (Auer, Dietzold et al. 2009)
- First departure: to avoid an overwhelming amount of triples they decided to split the storage of data between a database and a triple store. Lat/Lon data has been left stored in relational databases.
- Second departure: they elicited a class hierarchy, object properties and data attributes from OSM wiki page, this way they created the ontology also called the vocabulary.
- Extracted data.
- Developed a procedure to create *mappings* with DBpedia.
- Published data in the semantic web.
- Created an online browser-editor.

The above mentioned steps are in (Auer, Lehmann et al. 2009).

Created the project hosted in Google code.

Work is still ongoing, focusing more on data storage issues. Published data dates back the first issue in the summer of 2009.

### 4.4.2 The LinkedGeoData ontology

There is no reference to either cartographic or GI standard as in the original

---

<sup>58</sup> The “s” in childrens is in the cited text

<sup>59</sup> [http://wiki.openstreetmap.org/wiki/Machine-readable\\_Map\\_Feature\\_list#Use\\_cases](http://wiki.openstreetmap.org/wiki/Machine-readable_Map_Feature_list#Use_cases)

<sup>60</sup> Developed in Leipzig University by the same group developed DBpedia, the semantic translation of Wikipedia.

OSM project. LGD relied entirely on OSM every assumption derives from OSM structure. Therefore, the database structure is as follows.

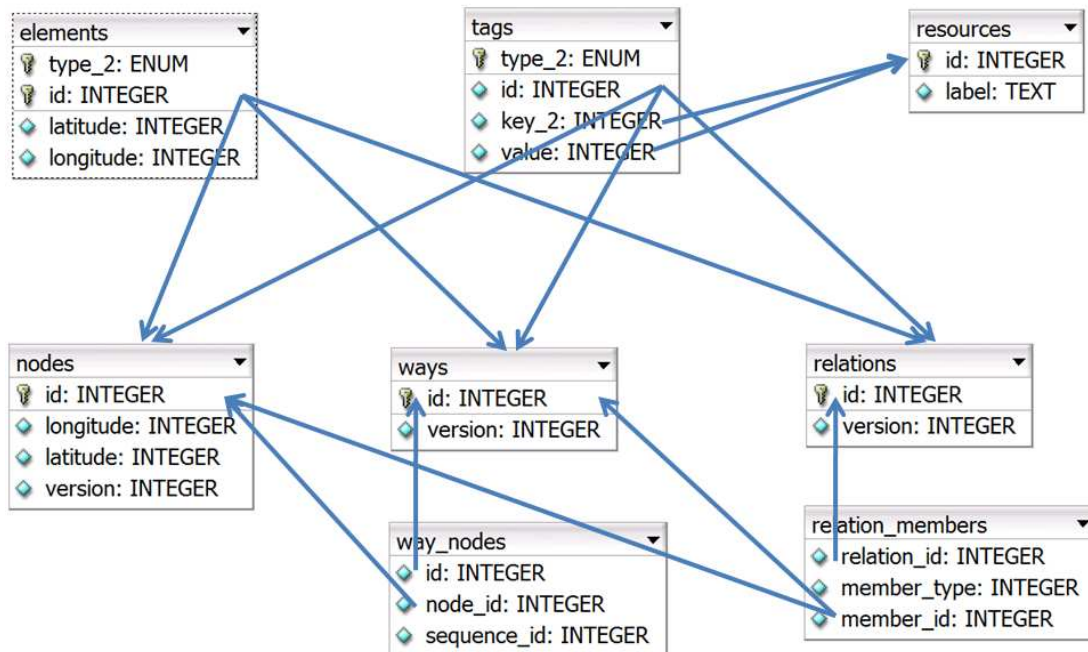


Figure: 4-1 LinkedGeoData database, from (Auer, et al. 2009, 1)

The ontology (the Tbox) downloaded from the project website has been opened using Protégé the ontology editor<sup>61</sup> and evaluated.

LGD has given a semantic structure to OSM data. The second above cited departure consisted in the recognition of three main families of tags:

- Classification attributes
- Description attributes
- Data attributes

Classification attributes led to the creation of class hierarchies, the couple key=value pair became a class/subclass pair.

Description attributes led to the creation of object properties and resources, the key=value pair became an objectProperty/resource pair.

Data attributes led to the creation of data type property therefore the key=value pair became a dataType/literals pair.

The upper level of the ontology (Tbox) is as in figure 4-2, following page:

<sup>61</sup> <http://protege.stanford.edu/>



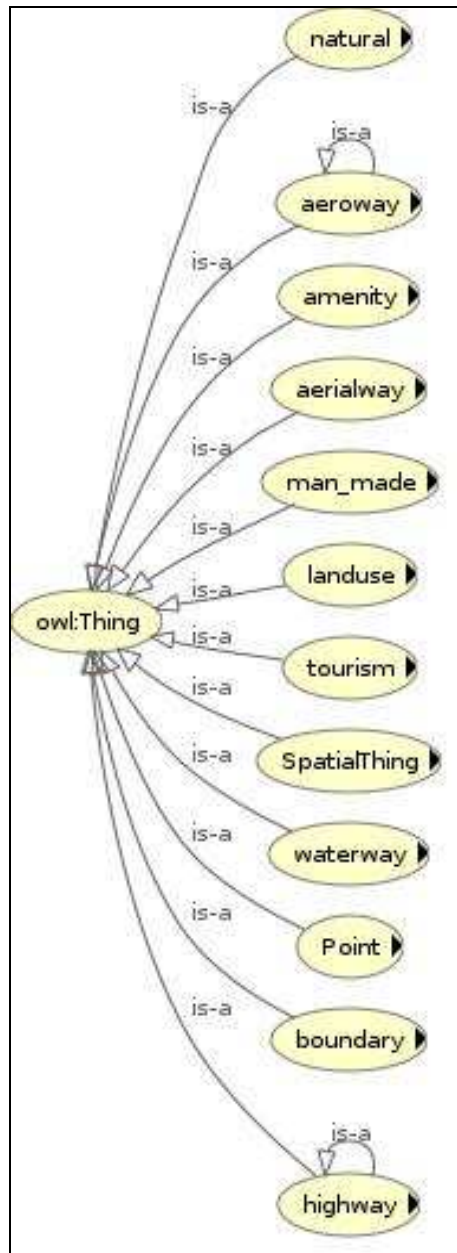


Figure: 4-2 LinkedGeoData upper level hierarchy

This is then the top level of classification attributes as chosen by LGD developers amongst the key=value pair of OSM. You can see a mixing of different abstraction levels coexisting at the same hierarchical level, both the data primitives and some tag values. Moreover, you can notice that both highway and aeroway are subclasses of themselves. Starting from the second level there is an overwhelming increase of classes, at the third level you can find millions of instances.

### 4.4.3 Ontology evaluation

To analyse the LinkedGeoData ontology due to the absence of metadata or standard evaluation methods for ontologies as outlined previously, Protégé metrics tab is used to better evaluate the ontologic expressiveness and the presence of the feature that characterizes a GI ontology as outlined in 4.1.

#### Ontologic precision

It is necessary to know whenever the LGD ontology falls into some of the identified categories in (Obrst, 2006). In 4.4.2 the procedure LGD developers used to create the ontology have been reported. Since they created a class hierarchy and stated object properties and data attributes, we can classify it like taxonomy. Some characteristics, however, like object and data properties afforded more expressivity to the ontology but they remain poorly structured. The OWL sublanguage used is OWL DL.

#### GI ontology features

Since the framework of (Tomai, Kavouras, 2004) in 4.1 is based on formal ontologies as outlined in 2.4.5, folksonomies do not fall into such a classification. In this section, a classification and an evaluation of the LGD ontology has been undertaken according to the classification schema previously explained. We have seen in 4.4.2 that LGD added to OSM formalization in RDF via the hierarchy of classes between mapped features that did not exist in the OSM database, object properties and data attributes. In the following page in Table 4.2 the evaluation of LinkedGeoData as a GI ontology according to the framework in *Table: 4.1 Geographic information ontology elements*. Therefore, the LGD ontology is missing almost all characteristics for a GI domain ontology. Its usage has to be carefully considered. OSM data need to be pre-processed before translated into ontology to find the hidden semantic relation between data.

GI ontology elements	Explanations and contextualization in GI	In LinkedGeoData
Concepts	Concepts involved in the scope of the ontology or coming from existing domain ontologies as outlined for domain ontologies above.	YES but strictly derived from OSM
Lexicon or vocabulary	Definition of concepts and relations in natural language plus metadata on how the ontology might be used.	YES but no metadata more than provided by original OSM data like author and changeset since the only objective of OSM and LGD is to create the database and to promote the project.
Relations as semantic relations	Relations between concepts at semantic lexical level: Synonymy, Hyperonym /hyponim, meronym/holonym they pertain to upper level ontologies.	Yes, we have the class hierarchy.
Relations as semantic property	Properties for concepts and values they can take according to the task of the ontology: Spatiality, temporality, Nature, Material/cover, Purpose, and Activity. Furthermore, the role means and results of the specific activity might be specified.	Spatiality is rarely present. Once the "is-in" key was used but it was not linking two concepts <sup>62</sup> , recently this kind of topological tagging has been deprecated.
Relations among relations	Taxonomy of relations between Semantic relations and semantic properties.	NO
Relations as axioms	Constraints on concepts and properties to assess consistency and completeness of the ontology: mereology, location, topology.	NO, allowed values are not formalized.

Table: 4.2 Geographic information ontology elements and LinkedGeoData

#### 4.4.4 LinkedGeoData mapping with DBpedia

LinkedGeoData has been published in Linking Open Data and has been mapped with DBpedia using a three criterion procedure as outlined in (Auer et al. 2009, 1)

The three criteria that have been used are as follows:

<sup>62</sup> The is-in relation we will see in data comparison doesn't involve the Ireland concept but only a string both in OSM and in LGD

- *Type information*
- *Spatial distance*
- *Name similarity*

It means that only entities in DBpedia that have lat-lon information have been included in the mapping. The mapping is at instance level since OWL DL is the employed OWL sublanguage.

To discover how many mappings have been stated and how many types of LinkedGeoData data have been mapped the SPARQL endpoint <http://lod.openlinksw.com/sparql> was queried. Moreover, the queries were directed to discover how many classes of the LGD ontology through their instances inherited a mapping with DBpedia. Therefore, the mapping triples from the LGD website were downloaded.

The triples are all in the form of S sameAs O. where S is the DBpedia instance and O is the LGD instance.

Following an example

<a href="http://dbpedia.org/resource/Pepsi_Center">http://dbpedia.org/resource/Pepsi_Center</a>	Subject
<a href="http://www.w3.org/2002/07/owl#sameAs">http://www.w3.org/2002/07/owl#sameAs</a>	Predicate
<a href="http://linkedgeo.org/triplify/way/25312645#id">http://linkedgeo.org/triplify/way/25312645#id</a>	Object

**Listing: 4-1 - Mapping triples between DBpedia and LinkedGeoData**

The following queries have been designed considering this structure of triples. Moreover, the mapping triples in the LOD are stored in the DBpedia graph so to construct queries involving the sameAs assertions we have to consider it. To deepen the knowledge of the mappings between LGD and DBpedia the SPARQL endpoint has been queried for all instances of LGD that have been mapped with DBpedia. The following query has been restricted to the LinkedGeoData graph:

```
select count distinct ?o where
{?s <http://www.w3.org/2002/07/owl#sameAs> ?o.}
```

**SPARQL listing 4.1 Mapped instances between DBpedia and LinkedGeoData**

We have 52.634 mappings between LGD and DBpedia. All types of mapped instances of LGD ontologies have been found through the following query restricted to [linkedgeo.org](http://linkedgeo.org) graph:

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
1. select count distinct ?ty where
2. {?o a ?ty.
3. ?s <http://www.w3.org/2002/07/owl#sameAs> ?o.
4. FILTER (?ty != <http://linkedgeo.org/vocabulary#way> && ?ty
!= <http://linkedgeo.org/vocabulary#node>)}
```

**SPARQL listing 4.2 Types of mapped instances between LinkedGeoData and DBpedia.**

Line 4 restricts the search excluding the OSM primitives node and way, which are unfortunately just another type of LGD data and can influence the queries.

It is necessary because with line 2 we can select all types of data related to mapped instances including the OSM primitives as can be seen in Listing: 4-3 *LinkedGeoData triples*.

Line 3 identifies all elements that own a mapping.

Line 2 identifies all types of LGD data with elements identified according to line 3 and 4.

Line 1 extracts all the types of data identified according to the constraints stated in lines 2 to 4.

Running the query 127 types of LGD data have been obtained, excluding the primitives node and way. Therefore, in the following sections when a line like the n<sup>2</sup> is used without the restriction of line 4, nodes and ways will increase the number of results exponentially and uselessly. The LGD types of data that owns at least one mapped instance with DBpedia are listed in Appendix.

We have 129 types of data involved in the mapping amongst 11.621 types of data in LGD ontology. Only 1,11% of data types owns a mapping.

Therefore, the mapping with DBpedia is at instance level and covers a very narrow part of LinkedGeoData. Classes are excluded from the mapping. Even though there are corresponding classes in LinkedGeoData and in DBpedia, they are not mapped. As an example DBpedia has the class "Hospital" <<http://dbpedia.org/resource/Hospital>> while in LGD we have the class "hospital" <<http://linkedgeo.org/vocabulary#hospital>> equivalence between those two classes would lead to interesting results but they are not mapped since OWL DL does not support such a statement.

#### 4.4.5 LinkedGeoData and OpenStreetMap - Taxonomies

Comparing LGD ontology with the upper level of the Map Features list<sup>63</sup> in OSM, which is a kind of taxonomy of OSM<sup>64</sup>, we can see that in OSM a first hierarchical level distinguishes between five kinds of data, excluding the data primitives. The taxonomy of OSM is only represented in the wiki page, LGD created its taxonomy choosing to differentiate the three families as depicted in 4.4.2. Therefore, the Tbox of LGD looks poorer if compared with the map features list but it is richer since it leads to information whereas the map features list is only a series of strings.

In the following page OSM map features list:

---

<sup>63</sup> [http://wiki.openstreetmap.org/wiki/Map\\_Features](http://wiki.openstreetmap.org/wiki/Map_Features)

<sup>64</sup> Cannot be considered as taxonomy of OSM because firstly, the classification is not in the data structure but only in the wiki, secondly the classification is only about some tags and users can use their own if they like. Taxonomies are prescriptive.

- |                 |                  |
|-----------------|------------------|
| 1 Physical      | 2 Non Physical   |
| 1.1 Highway     | 2.1 Route        |
| 1.2 Barrier     | 2.2 Boundary     |
| 1.3 Cycleway    | 2.3 Sport        |
| 1.4 Tracktype   | 2.4 Abutters     |
| 1.5 Waterway    | 2.5 Accessories  |
| 1.6 Railway     | 2.6 Properties   |
| 1.7 Aeroway     | 2.7 Restrictions |
| 1.8 Aerialway   |                  |
| 1.9 Power       | 3 Naming         |
| 1.10 Man Made   | 3.1 Name         |
| 1.11 Leisure    | 3.2 References   |
| 1.12 Amenity    | 3.3 Places       |
| 1.13 Office     | 3.4 Addresses    |
| 1.14 Shop       |                  |
| 1.15 Tourism    | 4 Annotation     |
| 1.16 Historic   |                  |
| 1.17 Landuse    | 5 Editor keys    |
| 1.18 Military   |                  |
| 1.19 Natural    |                  |
| 1.20 Geological |                  |

#### 4.4.6 LinkedGeoData and OpenStreetMap - Data comparison

Looking at an excerpt of data, we have chosen a node in OSM representing Dublin Airport and the triples in LGD that represent the same entity:

##### OpenStreetMap

```
<osm version="0.6" generator="OpenStreetMap server">
<node id="26608600" lat="53.4270037" lon="-6.243884" version="5"
changeset="3607418" user="Polarbear" uid="114161" visible="true"
timestamp="2010-01-13T01:15:14Z">
<tag k="name" v="Dublin International Airport"/>
<tag k="type" v="civil"/>
<tag k="iata" v="DUB"/>
<tag k="aeroway" v="aerodrome"/>
<tag k="is_in" v="Dublin,Ireland"/>
<tag k="icao" v="EIDW"/>
<tag k="source" v="Gagravarr_Airports"/>
</node>
</osm>
26608600>
```

Listing: 4-2 OpenStreetMap example data

##### LinkedGeoData

For brevity, has been used the following namespaces:  
lgdnode: <http://linkedgeodata.org/triplify/node/>

lgdvocab: <<http://linkedgedata.org/vocabulary>>  
rdfs: <<http://www.w3.org/2000/01/rdf-schema#>>

Different LOD resources have been used to create data in LGD.  
The license of data from creative commons, attribution from a domain in Purl.org, latitude and longitude data with projection draw on different domains.  
Type declarations are highlighted in magenta, there will be a reference to them in the following chapter.  
Metadata on origin of data, time and author of edit are also reported

```
lgdnode:26608600 rdfscomment "Generated by Triplify V0.6
(http://Triplify.org)" .
lgdnode:26608600 <http://creativecommons.org/ns#license>
<http://creativecommons.org/licenses/by-sa/2.0/> .
lgdnode:26608600 lgdvocab:#attribution "This data is derived from
information collected by the OpenStreetMap project
(http://www.openstreetmap.org)."
```

**lgdnode:26608600#id** <<http://purl.org/dc/elements/1.1/publisher>> "AKSW
research group (<http://aksw.org>)".

**lgdnode:26608600#id**
<[http://www.w3.org/2003/01/geo/wgs84\\_pos#long](http://www.w3.org/2003/01/geo/wgs84_pos#long)> "-
6.2439"^^<<http://www.w3.org/2001/XMLSchema#decimal>> .

**lgdnode:26608600#id** <[http://www.w3.org/2003/01/geo/wgs84\\_pos#lat](http://www.w3.org/2003/01/geo/wgs84_pos#lat)>
"53.4270"^^<<http://www.w3.org/2001/XMLSchema#decimal>> .

**lgdnode:26608600#id** <<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>> lgdvocab:#aerodrome> .

**lgdnode:26608600#id** lgdvocab:#iata "DUB" .

**lgdnode:26608600#id** lgdvocab:#icao "EIDW" .

**lgdnode:26608600#id lgdvocab:#is\_in "Dublin,Ireland"** .

**lgdnode:26608600#id** lgdvocab:#name "Dublin Airport" .

**lgdnode:26608600#id** lgdvocab:#source "Gagravarr\_Airports" .

**lgdnode:26608600#id** lgdvocab:#type "civil" .

**lgdnode:26608600#id** <<http://www.georss.org/georss/point>> "53.4270037
-6.2438840" .

<> <<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>>
<<http://example.org/provenance/ns#DataItem>> .

<> <<http://example.org/provenance/ns#createdBy>> \_:x2 .
\_:x2 <<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>>
<<http://example.org/provenance/ns#DataCreation>> .

\_:x2 <<http://example.org/provenance/ns#performedAt>> "2010-04-
16T07:36:11+02:00"^^<<http://www.w3.org/2001/XMLSchema#dateTime>>
.

\_:x2 <<http://example.org/provenance/ns#performedBy>> \_:x1 .

\_:x2 <<http://example.org/provenance/ns#usedData>> \_:x3 .

\_:x2 <<http://example.org/provenance/ns#usedGuideline>> \_:x4 .

\_:x3 <<http://example.org/provenance/ns#containedBy>> \_:x5 .

```
_:x5 <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>  
<http://example.org/provenance/ns#Document> .  
_:x5 <http://example.org/provenance/ns#retrievedBy> _:x6 .  
_:x6 <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>  
<http://example.org/provenance/ns#DataAccess> .  
_:x6 <http://example.org/provenance/ns#performedAt> "2010-04-  
16T07:36:11+02:00"^^<http://www.w3.org/2001/XMLSchema#dateTime>  
.  
_:x6 <http://example.org/provenance/ns#performedBy> _:x1 .  
_:x4 <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>  
<http://example.org/provenance/types#TriplifyMapping> .  
_:x1 <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>  
<http://example.org/provenance/types#DataCreatingService> .  
_:x1 rdfscomment> "Triplify V0.6 (http://Triplify.org)".
```

**Listing: 4-3 LinkedGeoData triples**

LGD itself therefore relies on different standardized resources online: the W3C standards and georss for geographic coordinates, the creativecommons for license, dublin core hosted by purl for metadata and registered domains in example.org for the origin. The topological part that involves a feature and string and not two features has been outlined in bold. The triplify tool, as explained in the introduction on LGD, created the attribution part and defined the type of aerodrome (aerodrome in a class in the LGD vocabulary/taxonomy) but improved OSM adding a hierarchical structure to data. Unfortunately, the automated creation of the hierarchy led to some misleading results, as we will see in *Chapter five: Improving OpenStreetMap retrieval using semantic technologies*.

## 4.5 SVG maps

Developed by Essecue, an OSM user, it is intended to improve rendering of information related to SVG maps created using Osmarender. It focuses on the rendering part of the project and uses the LinkedGeoData initiative for OSM data in RDF format. Looking at rendering targets it omits dealing with instances and focuses only on Tbox. The approach translates OSM data in RDF using the same primitives and developing the properties for visualization purposes. Hence object taxonomies and abstraction layering rely on LGD ones. This effort can be found at the following webpage: (<http://wiki.openstreetmap.org/wiki/User:Esscue/OsmarenderOWL>)

## 4.6 OpenStreetMap Wrapper

The OSM wrapper is an attempt by Simon Reinhardt to map not only OSM but also real world object identifiers in RDF. The wrapper compared to the other semantic initiatives for OSM is developed in a more independent way; it



treats OSM like a data provider hence distinguishing between the “original” OSM data characteristics and good features for geodata themselves. Reinhardt distinguishes between object in OSM and thing in real world, the only conscious attempt to differentiate abstraction levels that coexists in any GI database. The author outlines a procedure using an XSLT stylesheet to translate real world elements mapped in OSM into RDF representations of them. The ontology that the author refers to in the webpage describing the wrapper has still not been developed. The XSLT stylesheet is incomplete with several issues not yet resolved as you can read in the comments. The OSM wrapper can be found at the webpage (<http://osm.bloody-byte.net/documents/index.html>).

#### 4.7 Evaluation of semantic initiatives for OpenStreetMap

Following an evaluation table to compare the above-cited initiatives to use semantic technologies to improve or translate OSM.

Semantic effort	Domain	Scope	Stage	Mapping
Semantic mediaWiki (SMW)	Map feature list	Create map features list from descriptive page of tags	Taxonomy specified: Stage 2.2 as above cited	Internal and external with SMW
LinkedGeoData	OSM database	Semantic translation of OSM	Working, website and documentation available: Stage 5	DBpedia in Linked Open Data
SVG Maps	Osmarender	Improving rendering through contextual information added in SVG maps	Patch for SVG maps through Osmarender developed: Stage 2.2 completed	Using OSM elements of LGD
OSM Wrapper	OSM database	Mapping SW with real world objects in OSM	Designing XSLT: Stage 2.2 ongoing	Planned with GeoNames

Table: 4.3 Evaluation of ontologies designed for OpenStreetMap

In the fourth column, the maturity of the initiatives according to the schema in 2.8 has been evaluated.

Due to the incomplete development of all the other initiatives, only LGD amongst the above-cited initiatives can be considered a resource for the present thesis.

As the purpose of LGD is to translate a GI database and to share it in the SW, this is exactly what LGD stands for. We will see that the usability of LGD has to be carefully considered.

Usability is hindered by many factors:

- The way LGD taxonomy has been built and populated since we can see that the procedure created classes and populated them with the wrong instances.
- The chosen mapping with DBpedia, since if instances are wrongly classified you cannot rely on class hierarchy starting from mapped instances.
- The class hierarchy itself has some surprising properties (e.g. the classes that are subclasses of themselves as noted in 4.4.2).

## **PART II: THE DEVELOPMENT**

## 5 Chapter five: Improving OpenStreetMap retrieval using semantic technologies

### Introduction

In this chapter, the resource technologies and methodologies explained in the previous chapters will be used to target the research objectives. The semantic technologies and resources introduced in Chapter two will be used to overcome the limitations of geodata obtained through crowdsourcing as in the Problem statement in section 1.12.

Semantic technologies are defining the upcoming development of the web, the semantic web or web of data whose first bootstrapping initiative is the Linking Open Data initiative as explained in Chapter two. The Linking Open Data initiative (in section 2.5) is gathering semantically developed data on the web coming from all domains of human knowledge. Among all the fields of knowledge that are available on the LOD the linguistic resource WordNet RDF/OWL (introduced in section 2.6.3) is perfectly suited for the present work allowing for the investigation and extraction of linguistic relations between words. The linguistic resource will be used to improve the retrieval of OpenStreetMap collected geodata overcoming inconsistent tagging. Inconsistent tagging undermines the usability of data as mentioned in section 1.5. A semantic translation of OpenStreetMap is required to apply semantic technology. LinkedGeoData (introduced in section 4.4) will play this role. LinkedGeoData will be related to WordNet RDF/OWL in the web of data to make use of its linguistic potential. As we have seen in Chapter three both LinkedGeoData and WordNet RDF/OWL are in the LOD but they are not mapped directly.

In the present chapter will be investigating the possibility of coupling LinkedGeoData and WordNet RDF/OWL. The coupling will rely on the design of SPARQL queries that will drive the development of knowledge, linguistically expanding a requested keyword to its synonyms. SPARQL queries will be designed to target all LinkedGeoData elements that are associated with the queried keyword. This way linguistically inconsistent tagging in OpenStreetMap can be overcome. Since LinkedGeoData and WordNet are not mapped directly, we can follow two paths to couple them. The first one is based on a graph pattern (developed in section 5.4). Semantic data and relationships are expressed through triples as explained in Chapter two (section 2.2) with reasoning following patterns cascading from different knowledge areas. A first query then will be designed to follow a graph pattern along semantic relationships in the LOD. This query relies on every single semantic resource and on every single resource mapping encountered along the path. Since the information flow will also be directed along a path it will not

be very resource demanding.

Semantic resources on the LOD can also be coupled to overcome patterns. Since data can be discovered through filtering functions, a second query will be designed to connect directly LinkedGeoData and WordNet RDF/OWL restricting the discovery to the graphs pertaining to the resources to be coupled. The second query (developed along section 5.6) will rely only on the two identified resources; moreover, it is not related to the inner structure of resources as it jumps directly from the outcomes of WordNet RDF/OWL to LinkedGeoData resources. This kind of query having to browse in the LOD is considerably more resource demanding than the first one.

At the end of the chapter, an overall evaluation of the results achieved in the chapter (in section 5.7) includes an envisaged application for the “core component” here developed.

## **5.1 Use case, constraints and advantages of the chosen approach**

The use case is the following.

Someone relying on a mobile device wants to query the OSM database to find a location. If this person poses his or her request directly to unprocessed OSM data, due to the freedom of tagging embedded in OSM project, he or she would probably obtain limited answers because without strict guidance OSM members can use different synonymous words that describe the same real world objects or human activities. Therefore, if the requester uses a less widespread keyword he would probably obtain results restricted to the word. While it is positive to empower users, allowing them to add data without strict requisites, this is now limiting the usability of the dataset. Data consumers that have different linguistic tendencies would have limited results to queries. Moreover, in a broader perspective the generic geoinformation data consumer is not aware of standards on nouns endorsed in the OSM community. Therefore, queries could be formulated on neither documented nor used nouns. Hence, usability of data collected in the OSM project for the mobile application is twice hindered. The linguistic enrichment in queries has been introduced to overcome the limitation of the web 2.0 built database and to improve its usability via a generic geodata information requester through a mobile device.

### **5.1.1 The constraints**

To have simplicity, usability, portability and extensibility of the present work to other semantic resources available on the LOD some constraints have been stated. Those constraints might be easy to follow due to the semantic nature of used data:

The requester has to ask just a simple keyword and obtain the resulted geodata.

Since semantic is embedded in resources (at least the lexical ones), the

present work has been developed:

- To avoid any intermediate interaction with the end user.
- To avoid programming.

Since the requester is using a mobile device, he will query the LOD through SPARQL queries conveyed by HTTP requests, without requiring a local copy of the semantic Linked Open Data resources.

### **5.1.2 The advantages**

To take advantage of the published data in the Linked Open Data, SPARQL queries have been designed. Queries through the linguistic resource WordNet are able to identify not only the features corresponding to the requested term but also features that are tagged with synonymous terms. In further research, other linguistic relations between nouns (as in section 2.6.1) could also be implemented.

The query will rely on open published data so the requester will only need internet access and a GUI to insert the requested term (or keyword) in the queries. To achieve such a result queries have been run using the LOD SPARQL endpoint (<http://lod.openlinksw.com/sparql>). Queries can be made in several ways and, due to the enormous amount of data in the LOD, links existing among resources have been investigated. The project is not a standalone application but a bridge to a broader perspective of knowledge integration. It is an exploratory project since the data that has been coupled is only a very small fraction of the knowledge in the semantic web. Any kind of human knowledge that has been inserted in the semantic web has the potential to be coupled with geodata.

### **5.1.3 The procedure**

The queried keywords have been selected among OSM tags identifying classes of real world geo objects or properties associated with them. Two different approaches have been tested in constructing queries to integrate LGD and WordNet. Once again, we have seen in (Tomai & Kavouras, 2004) as outlined previously in section 4.1 *Geographic Information ontologies* that a Geo ontology might entail relationships between features that are semantic relations like synonymy, hyperonymy/hyponymy or meronymy/holonymy. The LinkedGeoData ontology is totally missing semantic relations among geographical objects. In the present work, this limit will be overcome, working only on synonymy, coupling LinkedGeoData with a linguistic resource. A geo ontology will be indirectly integrated with lexical information working only on the synonymy aspect. A slightly more complex approach can easily expand to other semantic relations in geodata. The present thesis is a starting point to evaluate the potential and the features geoinformation might have to be fully integrated and accessible in unprecedented and unpredicted ways in the semantic web. It was one of the purposes of the openness of both the semantic web, in its core component of the Linking Open Data initiative, and OpenStreetMap.

The queries will be constructed to answer a statement such as the following:

*Where I can find a bakery?*

Since the word bakery has two synonymous (bakeshop and bakehouse) we are not sure that all OSM users tagged all the workplaces where *baked goods (breads, cakes and pastries) are produced or sold*<sup>65</sup> using only one of those nouns. To target all possible synonyms users might have used, the query has to be expanded to include all terms related through synonymy to bakery. Although our reasoning creates associations between objects creating classes (bakeries, groceries) the present work has been carried out using links between every individual object represented by instances (bakery1, bakery2). This constraint comes from the ontologies involved that have been developed using OWL DL. As explained in Chapter three: LinkedGeoData, WordNet RDF/OWL and DBpedia have been developed using OWL DL; therefore, the sameAs assertions have been used to create mappings between instances and not between classes. The OWL sublanguage 'OWL DL' does not allow sameAs assertions between classes. This kind of assertion is possible only using OWL Full. The procedure to obtain the queried locations will follow the three steps below:

*like this* Step: find words linguistically related to the queried term through WordNet RDF/OWL.

*places* Step: find in LGD classes of instances that contains not only the features/instances representing the queried term but also the features/instances linguistically related to it as found in the first step.

*where* Step: obtain the URI representing the searched spatial entities.

In the present work, geographic entities like city, province and state will be not investigated because the web has plenty of applications able to target this result. Although the integration between LinkedGeoData and WordNet could be beneficial for LinkedGeoData, since WordNet relates geographical entities semantically, WordNet is not a gazetteer therefore lacking a lot of locations or geographical entities<sup>66</sup>. Moreover, the present work focuses on the usability of peer produced geoinformation that can be affected by inconsistent tagging, integrating local knowledge in the semantic web. The present work has been conducted designing queries over the LOD. The queries have been tested over a number of keywords that have been chosen amongst tags in OSM. Tags have been translated in classes or datatype properties by LinkedGeoData developers as in 4.4.2. In the next section, the selection of the sample keywords shall be explained.

---

<sup>65</sup> Definition of bakery and its synonymous taken from WordNet3.0  
<http://wordnetweb.princeton.edu/perl/webwn?s=bakery&sub=Search+WordNet&o2=&o0=1&o7=&o5=&o1=1&o6=&o4=&o3=&h=>

<sup>66</sup> E.g. in Sicily amongst nine provinces only four are in WordNet 3.0. They are semantically related to Sicily by a *part meronym* relation

## 5.2 The sample

Since in OSM a tag is a key=value pair, where the key is a more general term and the value a specification of one property of the key. The sample has been selected trying to integrate some interesting values of shop=\* coming from the subclasses of “shop” in the LGD ontology, with some very commonly used tags that aren’t documented as keys in the OSM's map features list. The list is the main reference for OSM mappers collecting all shared definitions of geographical objects and the suggested tags or combinations of them. The map feature list is the weak standardization effort of the OpenStreetMap community. The tagwatch<sup>67</sup> service has been used to find undocumented but commonly used tags. The classes “grocery” “bakery” and “technology” have been chosen amongst the subclasses of the “shop” class in LGD ontology. Bakery is a documented key while grocery and technology are undocumented keys. The classes “path” and “footway” have been added due to a long debate in OSM community concerning the distinction between those two terms both path and footway are documented tags. Most undocumented keys coming from four countries have been selected. According to tagwatch statistics, Germany is the European country with more used tags. In Great Britain OSM mappers, using their native language for tags, will have a wider linguistic spectrum than all the other countries. Netherlands had a great improvement in OSM database by massive publicly owned dataset imports; while in Italy the import of publicly owned datasets is very rare and tagging comes primarily from users’ contributions. Amongst the 100 top undocumented keys in the above mentioned countries the following in Table: 4-1 have been chosen:

Germany	Netherlands	Great Britain	Italy
parking*	technology	site	maintenance
shelter*	frequency*	direction*	gritting
Network			
footway*			
hiking*			

**Table: 5.1 Sample keywords from national statistics**

They are material and immaterial ones, classes of objects and object properties, ranging from shelter to frequency. Few of them are really undocumented since they are not documented as keys but they are used as values. (E.g. parking is documented as a value in the couple amenity=parking, the usage of a key parking=\* is undocumented). Keys with asterisks are documented only as values. Undocumented tags are preferred thus to enrich the retrieval of geographic information. The tag “stadium” has been added because it is surely not only documented but also easy to find in any knowledge base on the LOD, it will help in testing safer paths. Summing up the list of the sample keywords in the following table:

<sup>67</sup> <http://tagwatch.stoecker.eu/>



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

**Table: 5.2 - Sample keywords**

To achieve background information on selected keywords a SPARQL query had to be designed to determine the classification and the type of data of those tags in LGD. The UNION construct that allows a SPARQL query to have alternative results has been used thus:

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX lgdvocab: <http://linkededgeodata.org/vocabulary#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
select distinct ?Super ?type ?Sub where
  1. {{lgdvocab:technology a ?type.}union
  2. {lgdvocab:technology rdfs:subClassOf ?Super.} union
  3. {?Sub rdfs:subClassOf lgdvocab:technology. }}
```

**SPARQL listing 5.1 sample keywords in LGD**

In line one we ask for the type of data technology pertaining to (class literal or property).

In line two we ask for the superclasses of the class the keyword belongs to.

In line three we ask for the subclasses of the class the keyword belongs to.

In Table 5.3, following page the results for the query repeated for every keyword are shown.

Sample	Type	Superclasses	Subclasses
bakery*		lgdvocab:shop	
direction*	owl:DatatypeProperty		
footway*	owl:DatatypeProperty owl:Class	lgdvocab:highway	
frequency*	owl:DatatypeProperty		
gritting	owl:DatatypeProperty		
grocery	owl:Class	lgdvocab:shop	
hiking*	owl:DatatypeProperty owl:Class	lgdvocab:route	
maintenance	owl:DatatypeProperty		
network	owl:DatatypeProperty		
parking*	owl:DatatypeProperty	lgdvocab:amenity lgdvocab:tourism lgdvocab:building lgdvocab:landuse	
path*	owl:DatatypeProperty owl:Class	lgdvocab:highway lgdvocab:cycleway lgdvocab:route	
shelter*	owl:DatatypeProperty	lgdvocab:amenity lgdvocab:building	
site	owl:DatatypeProperty		
stadium*	owl:Class	lgdvocab:leisure lgdvocab:building	
technology	owl:DatatypeProperty owl:Class	lgdvocab:shop	

**Table: 5.3 Sample keywords: types, subclasses superclasses**

In the first column, you have the queried keywords. In the second column, the types of data as stored in the variable ?type in line1 are listed. In the third column the superclasses in LGD ontology for the queried keyword as stored in the ?Super variable in line 2. In the fourth column, the values stored in line 3 for the variable ?Sub that stored all subclasses for the given keywords.

No subclasses have been found.

The sample offers a wide spectrum of possibilities to test the approach since Classes and Datatype properties have been selected from different levels of LGD taxonomy. All classes have no subclasses so they contain only instances. Since the mapping between LGD and DBpedia is at the instance level, information flow has to be followed at the instance level. Classes will be used to underline classifications in the different ontologies and an attempt to let them inherit mappings from instances will be explained.

### 5.3 Matching semantic resources on the LOD

Ontology matching has been introduced in 2.4.2. Statistics on the matching between WordNet RDF/OWL and DBpedia has been published for an OAEI initiative (OAEI, 2009). The matching of pairs of ontologies at a time will be developed in the following sections. To enrich the query patterns that have been found through mapped instances<sup>68</sup> in the LOD, a word will be connected geo objects. Broadening matches from instance to class level will be performed to gain better results. The present work tests how the LOD can be used to run over complex queries that embed geodata and different sources of knowledge. Ontology matching has been used to match different datasets as in (Buccella et al., 2009) while the use of WordNet to expand queries has never relied on the RDF/OWL version of it and consequently neither on the web of data. The task of the present work is to find a way to query the web of data more easily finding, if possible, the links between the resources LGD and WordNet RDF/OWL. If we look at the cloud representation of the Linking Open Data as in Figure 5-1 below we see that the two datasets we want to work together are not mapped directly since there is not an arrow that links them.

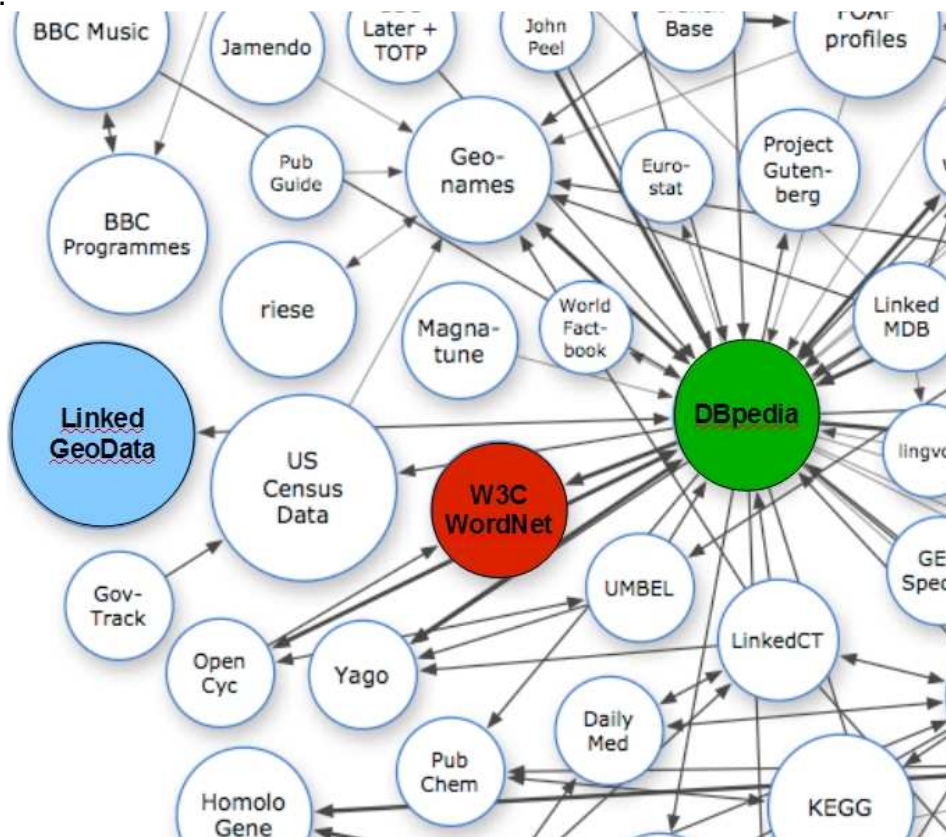


Figure: 5-1 Involved resources on the LOD cloud.

<sup>68</sup> As outlined previously it is compulsory working at instance level because all matching are at instance level and the sameAs assertions in OWL DL are not allowed between classes.

The light blue LinkedGeoData circle and the red W3C WordNet circle are both mapped with the green DBpedia circle but there is not an arrow directly between them. LinkedGeoData and WordNet RDF/OWL are both mapped with DBpedia and therefore the shortest path to explore is passing through DBpedia. Other ways might be found using semantic web search engines like the ones listed in (Bizer Heath Berners-Lee, 2009), since other ontologies linked with DBpedia own mappings with WordNet RDF/OWL In this work, other ways have been avoided since LGD is mapped only with DBpedia. Browsing SW portals, you can find lots of SW search engines and browsers. The reliability of some services is dubious<sup>69</sup>. In the present work, the instances of “stadium” class and “shop” class and its subclasses in LGD and their lexical form in WordNet RDF/OWL have been investigated to test the path between resources. Two methodologies have been explored to design SPARQL queries to enrich the retrieval of geodata. The first is based on LOD surfing and will be *query1: graph pattern query* in section 5.4. A graph pattern that connects to WordNet RDF/OWL from LinkedGeoData via DBpedia has been found and evaluated. The query will be directed along a path linking instances and datatype properties starting from the queried term and finding at the end of the pattern the requested geographical entities mapped initially by OSM users. Having to cross DBpedia, the success of this query has been related to the influence of the three ontologies involved. This pattern is the one that might be covered in the opposite direction by the information flow originated with the insertion of the queried term. The second query will be designed to link directly WordNet RDF/OWL and LinkedGeoData through string matching forming *query2: string matching query* developed in section 5.6. It will not rely on graph patterns from one resource to the other and so its performance will be not limited by everything that is between them in the graph pattern query, namely the two mappings and the DBpedia resource. It will rely only on the originating and target ontology. Moreover, the string match query will directly target classes in LGD ontology and will not rely on instance level matching.

## 5.4 Query1 - The graph pattern

In the following sections, the making of the graph pattern query will be explained as represented by *Figure: 5-2 Query1 - Graph pattern blocks*, following page. The boxes representing the resources while the thick double arrows representing the mappings between the resources.

---

<sup>69</sup> Using The Disco Hyperdata browser ([http://www4.wiwiwss.fu-berlin.de/rdf\\_browser/](http://www4.wiwiwss.fu-berlin.de/rdf_browser/)) we have to go at instance level find that The shop class in the LGD ontology is a subclassOf... itself and surprisingly is a subclass of the vocabulary directly but also a subclass of all the shop classes.(?) It looks surprisingly because since the class shop in the LGD ontology is the superclass of all the shops. It means that the tool really meant that there is a subclassOf relation between shop and all shop type but looks like the tool is not able to visualize the distinction between subjects and objects of triples.



Figure: 5-2 Query1 - Graph pattern blocks

The double arrows are thick since they are triple patterns, they are overlapping resources since the mapping is based on semantic relations, the arrows are thinner than boxes because not all resources are mapped.

During an OAEI campaign (OAEI, 2009) it has been estimated that only 35% of DBpedia is mapped with WordNet RDF/OWL. A brief analysis of the inner structure of the three resources of the LOD and their mappings has been conducted to find a graph pattern that from a simple lexical form can drive the extraction, through a SPARQL query, all the objects or activities on the Earth's surface linguistically related to the lexical form by synonymy. The objects or activities on the Earth's surface are the ones surveyed by OSM users and translated into semantic resources by the LinkedGeoData initiative.

Due to the enormous amount of data, the analysis has been restricted to finding the existing possibilities aimed at discovering the instances of the class "stadium" and subclasses of "shop" class in LGD. The resources in the boxes have been described in section 2.5. *The environment of the proposed development: Linking Open Data*. In the following sections, the parts involved in the mapping will be explored. Firstly the starting point of the LOD surfing will be analysed, that is LGD (the cyan box on the right in Figure: 5-2) focusing on mapped instances between LGD and DBpedia (the orange double arrow on the right in Figure: 5-2). Since the mapping between DBpedia and WordNet RDF/OWL is stated through property, the investigation will follow the mapping between mapped instances and WordNet RDF/OWL (the yellow double arrow on the left in Figure: 5-2). Finally in WordNet RDF/OWL (the red box on the left in Figure: 5-2) the extraction of synonymous given a word will be shown.

In the Figure: 5-3 Query1 Graph pattern at instance level, following page, a more detailed description has been designed. The big cyan arrow on the left expresses the direction of our query building, while at runtime the query will act in the opposite direction starting from the magenta rhombus to drive the requester to the blue LGD circles. In the following sections is an explanation of the first mapping, the orange one on the right in *Figure: 5-2* between LinkedGeoData and DBpedia. The path will be designed starting from instances of LGD that belong to subclasses of the "shop" class and instances of the "stadium" class (the blue circles in Figure 5-3) that are related to DBpedia instances (light orange circles in Figure 5-3) through a sameAs assertion as designed in Figure: 5-3 Query1 Graph pattern at instance level.

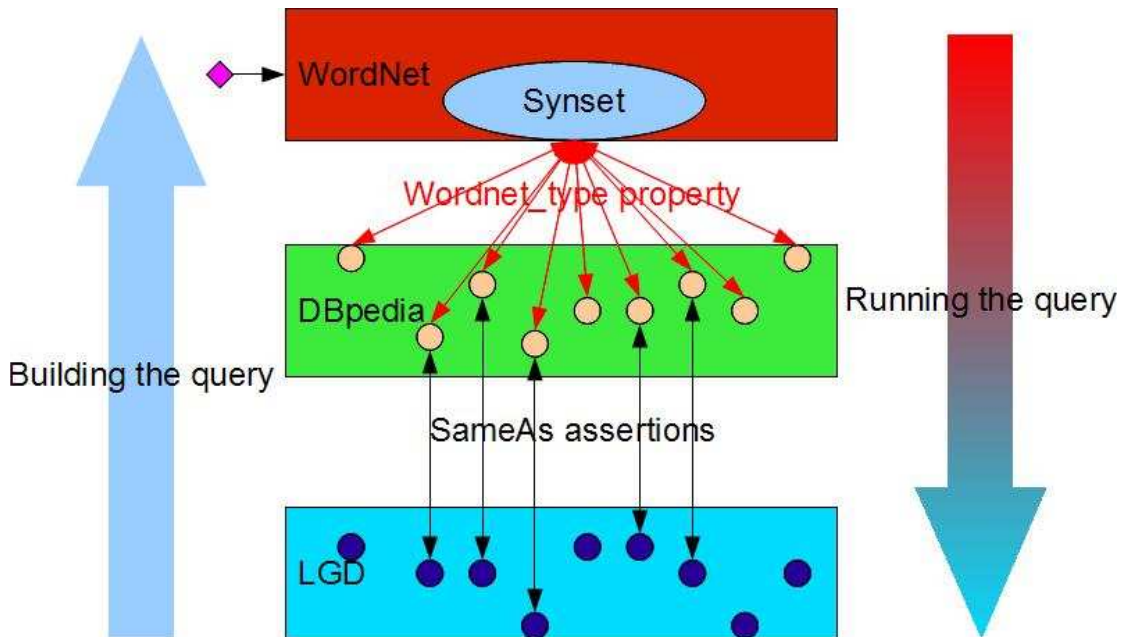


Figure: 5-3 Query1 Graph pattern at instance level

#### 5.4.1 LinkedGeoData mappings with DBpedia

First analyses on the mappings between LinkedGeoData and DBpedia have already been shown in 4.4.4. It has been found that there are 52.634 mappings between LGD and DBpedia. Those instances pertain to 129 types of LGD data including the primitives “node” and “way”. The LGD type of data that owns at least one mapped instance with DBpedia are listed in Appendix. We have 129 types of data involved in the mapping amongst 11.621 types of data in LGD ontology. Only 1% of data types own a mapping. It constitutes a considerable bottleneck to query performance using a graph pattern. Since we found a mapping between instances of classes and properties, a method has been tried to expand the link obtained through mapping to all instances of the involved classes and properties. Considering the cyan LinkedGeoData box in Figure: 5-3 Query1 Graph pattern at instance level we see that the mapping involved some, but not all, instances of a given class. Unfortunately, as we have seen in the magenta triples in Listing: 4-3 *LinkedGeoData triples* the type declarations in LinkedGeoData involves many different types of data for any purpose. Therefore using the following SPARQL query:

```

1. select COUNT DISTINCT ?t where
2.  {?s <http://www.w3.org/2002/07/owl#sameAs> ?o.
3.  ?o a ?ty.
4.  ?t a ?ty.
5.  FILTER (?ty != <http://linkedgeodata.org/vocabulary#way> &&
           ?ty != <http://linkedgeodata.org/vocabulary#node>)}
    
```

SPARQL listing 5.2 Type expansion for query1 graph pattern

We obtain too many instances to be computed overcoming the technological

limit of 2.097.151 rows for the SPARQL endpoint. The mapping has been expanded by more than 4000% but also includes many useless triples. In the query above in line 2 “?o” the variable that collects all mapped entities is found. With the variable “?ty” all the type of data that have amongst them mapped entities have been identified. Therefore in line 4 in “?t” all the instances of “?ty” type of data have been identified thus including mapped and unmapped entities and all other type declarations that we found in *Listing: 4-3 LinkedGeoData triples*. In the foregoing query, a kind of **double instantiation assertion** has been used, in line 3 and 4. The double instantiation assertion has been used several times in the present work. It is a useful construction for the expansion of a relation from an instance to all types of data of the same kind. In the foregoing query *SPARQL listing 5.2 Type expansion for query1 graph pattern* the attempted expansion involves the link represented by the mapping between LGD and DBpedia.

Going back to our sample, we can now see if the entities pertain to linked types. The Table 5-4 on the following page contains the sample and the mapped types. In table 5.4, the first column lists the sample keywords, the second column lists the datatype and the third column witnesses the existence of mapping for some element of the same type. Only one keyword of our sample owns some instances of its mapped class. One out of fifteen or 6,7%, moreover “stadium” is the only keyword that has been added with the deliberate purpose of testing the methodology using a commonly used word.

Sample keywords	type	Mapped
bakery*		No
direction*	owl:DatatypeProperty	No
footway*	owl:DatatypeProperty owl:Class	No
frequency*	owl:DatatypeProperty	No
gritting	owl:DatatypeProperty	No
grocery	owl:Class	No
hiking*	owl:DatatypeProperty owl:Class	No
maintenance	owl:DatatypeProperty	No
network	owl:DatatypeProperty	No
parking*	owl:DatatypeProperty	No
path*	owl:DatatypeProperty owl:Class	No
shelter*	owl:DatatypeProperty	No
site	owl:DatatypeProperty	Part of some name
stadium*	owl:Class	YES
technology	owl:DatatypeProperty owl:Class	No

Table: 5.4 Sample keywords and mapping between LinkedGeoData and DBpedia

Running the following query we want to know how many stadiums are mapped with DBpedia and how many stadiums we have in the LGD class:

```
select COUNT DISTINCT ?t where
  1. {?s <http://www.w3.org/2002/07/owl#sameAs> ?o.
  2. ?o a <http://linkedgeoata.org/vocabulary#stadium>.
  3. ?t a <http://linkedgeoata.org/vocabulary#stadium>.
```

SPARQL listing 5.3 Mapped and unmapped "stadium" in LinkedGeoData

In the first line all mapped instances have been selected where the variable ?o collects all instances from the LinkedGeoData side, in the second line the variable ?o restricts its values to the elements that are of type "stadium" in LinkedGeoData. In line three all instances of "stadium" LGD class have been selected in the variable ?t. The variable ?o gives 345 mapped stadiums over 5.244 stadiums in LGD recognized through variable ?t. So only 6,58% of stadiums in LGD are mapped with DBpedia. The graph pattern between LinkedGeoData and DBpedia has to follow the mappings, unfortunately only stadium class amongst all keyword in the sample owns mappings. Following the mapping between DBpedia and WordNet RDF/OWL.

#### 5.4.2 DBpedia mappings with WordNet RDF/OWL

The last mapping between semantic web resources that was investigated is the one between DBpedia and WordNet RDF/OWL. The starting point is the set of DBpedia instances that are mapped with instances of the "stadium" class in LGD. The alignment with DBpedia is at the instance level. The mapping explored here is represented in *Figure: 5-2 Query1 - Graph pattern blocks* with a double thick yellow arrow. As anticipated in *Figure: 5-3 Query1 Graph pattern at instance level* the mapping between those two resources is achieved through a property: WordNet\_type.

DBpedia instances are related to synsets of WordNet RDF/OWL through triples whose subject is the feature, the predicate is the wordnet\_type property, and object is the synset describing the type of object in WordNet. Following is an example of the link between DBpedia and WordNet in the LOD.

```
<http://dbpedia.org/resource/Colis%C3%A9e_Pepsi>
<http://dbpedia.org/property/wordnet_type>
<http://www.w3.org/2006/03/wn/wn20/instances/synset-stadium-noun-1>.
```

Listing: 5-1 Mapping example between DBpedia and WordNet RDF/OWL

A graph pattern that drives from WordNet synsets to LGD resources has been found. The mapping holds for one only of the sample keywords.

The path from LGD resources and the keyword posted from the requester has followed the cyan arrow in *Figure: 5-3 Query1 Graph pattern at instance level*. The Pattern from LGD instances (the blue circles in *Figure 5-3*) through the sameAs assertions reached DBpedia instances (orange circles in *Figure 5-3*), this time from DBpedia instances through the wordnet\_type property reaching a synset in WordNet RDF/OWL.

The following section will explain how synonymous words can be found in



WordNet RDF/OWL starting from a given keyword. The following section will explain what happens inside the red box of Figure 5-2 and 5-3.

### 5.4.3 Extracting synonyms in WordNet RDF/OWL

Due to the structure of WordNet RDF/OWL, as explained in 2.6.3, it is easy to create a query to extract synonyms of a given word. The query is a graph pattern made by a sequence of six triples patterns. The same query can be run for every keyword of our sample:

```
PREFIX wn20schema: <http://www.w3.org/2006/03/wn/wn20/schema/>
SELECT distinct ?lexforms WHERE
1. {?bWords wn20schema:lexicalForm ?lexforms .
2. ?bWordSense wn20schema:word ?bWords .
3. ?aSynset wn20schema:containsWordSense ?bWordSense .
4. ?aSynset wn20schema:containsWordSense ?aWordSense .
5. ?aWordSense wn20schema:word ?aWord .
6. ?aWord wn20schema:lexicalForm "bakery".}
```

SPARQL listing 5.4 Extracting synonymous words of bakery in WordNet RDF/OWL

Similar queries can be easily designed for the other semantic relations between nouns. Two triple patterns at line 3 and 4 are a double instantiation to extract from the ?aSynset variable all synonymous words. In *Figure: 5-4 Extracting synonyms using WordNet RDF/OWL*, is a graphical representation of the graph pattern.

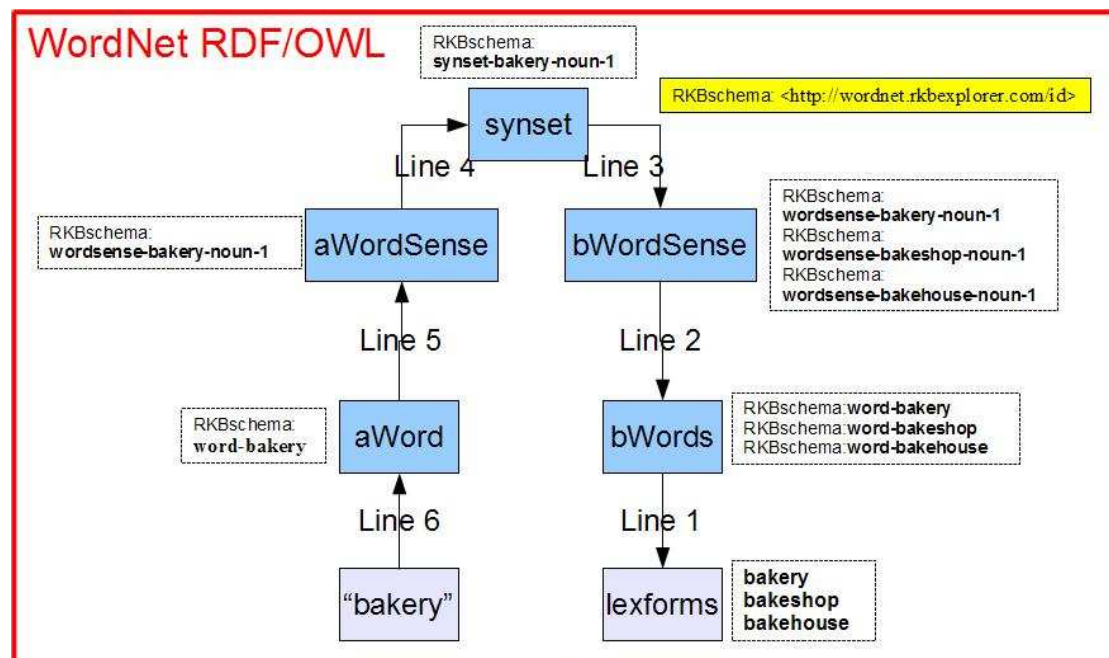


Figure: 5-4 Extracting synonyms using WordNet RDF/OWL

Every predicate is represented by an arrow, every triple pattern is identified

with the number of the corresponding line, subjects and objects are represented by cyan boxes. In dashed boxes, the values for every variable given the keyword bakery are reported. You can follow line by line the path of information driven by the SPARQL query above both in the Figure 5-4 and in the *Table: 5.5 Stages and variable values of bakery synonym extraction using WordNet RDF/OWL*.

In the graphical representation the information path goes up from a lexical form to a word to a wordsense to the synset, and then goes down from the synset to all the words contained in the synset, thus through the wordsense then the words and finally the lexical forms. Looking at the figure, we can follow the flux of information and the values of the variable line by line.

In the following tabular representation the values of variables achieved line by line (triple pattern by triple pattern then).

Line	Variables	Values
6	?aWord	<a href="http://wordnet.rkbexplorer.com/id/word-bakery">http://wordnet.rkbexplorer.com/id/word-bakery</a>
5	?aWordSense	<a href="http://wordnet.rkbexplorer.com/id/wordsense-bakery-noun-1">http://wordnet.rkbexplorer.com/id/wordsense-bakery-noun-1</a>
4	?aSynset	<a href="http://wordnet.rkbexplorer.com/id/synset-bakery-noun-1">http://wordnet.rkbexplorer.com/id/synset-bakery-noun-1</a>
3	?bWordSense	<a href="http://wordnet.rkbexplorer.com/id/wordsense-bakery-noun-1">http://wordnet.rkbexplorer.com/id/wordsense-bakery-noun-1</a> <a href="http://wordnet.rkbexplorer.com/id/wordsense-bakeshop-noun-1">http://wordnet.rkbexplorer.com/id/wordsense-bakeshop-noun-1</a> <a href="http://wordnet.rkbexplorer.com/id/wordsense-bakehouse-noun-1">http://wordnet.rkbexplorer.com/id/wordsense-bakehouse-noun-1</a>
2	?bWords	<a href="http://wordnet.rkbexplorer.com/id/word-bakery">http://wordnet.rkbexplorer.com/id/word-bakery</a> <a href="http://wordnet.rkbexplorer.com/id/word-bakeshop">http://wordnet.rkbexplorer.com/id/word-bakeshop</a> <a href="http://wordnet.rkbexplorer.com/id/word-bakehouse">http://wordnet.rkbexplorer.com/id/word-bakehouse</a>
1	?lexforms	<b>bakery</b> <b>bakeshop</b> <b>bakehouse</b>

**Table: 5.5 Stages and variable values of bakery synonym extraction using WordNet RDF/OWL**

In line 6 (first tuple), the queried term “bakery” in WordNet among words (bakery box) is identified.

In line 5, the Wordsense related to the word has been identified.

Line 4 and 3 is the double instantiation declaration you enter with one wordsense (?aWordSense) and extract all word senses (?bWordSense) contained in the synset as can be seen in the values of the variables in the dashed boxes.

In line 4, the synset that contains the given wordsense is identified.

In line 3, all other wordsenses included in the synset found in line 4 are extracted. Notice that from this point we will have three values for the variables since now our variables embed the three word senses that were included in the synset.

In line 2, the words related to the wordsenses found in line 4 are identified

In line 1, the lexical forms representing the words identified in line 2 have been found. In the following *Table: 5.5 Stages and variable values of bakery synonym extraction using WordNet RDF/OWL*, The URIs assigned to the variables along the graph pattern line by line. As outlined previously, passing from line 3 to line 4 we have a triple of values for our variables having extracted from the synset the three wordsenses including the one that in the procedure led us to the synset. In the following a query has been created from all the parts of the query investigated above that, through a graph pattern from the queried term, gives us all the entities in LGD ontology that are related to it linguistically through synonymy.

## 5.5 Query1 Graph pattern runtime

In the previous sections all the information regarding the mappings between the resources that will be involved in the design of the first query have been collected. The first query will be directed along semantically related information in the semantic web. The full query is built using a pattern of nine triples. Therefore, in this section the full pattern will be explained incrementally adding lines and explaining their role. Referring to Figure: 5-3 Query1 Graph pattern at instance level, in this section we are starting from the top inserting our keyword "stadium" in the red box WordNet. To better explain and measure the performance of the query it has been divided into smaller parts and the partial results of the query will be shown. The first part of the graph pattern query from line 1 to line 5 is represented in *SPARQL listing 5.5: Query1 graph pattern from WordNet to DBpedia* below. Focusing on the linguistic part of the query from line 1 to line 4; line 4 has been added due to the WordNet storage in the LOD, otherwise the same synset would have two different URIs and the mapping between WordNet and DBpedia would fail. It was not necessary to extract all synonyms of the given lexical form due to the fact that the mapping between WordNet and DBpedia is at synset level and not at instance level.

```
PREFIX wn20schema: <http://www.w3.org/2006/03/wn/wn20/schema/>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX dbprop: <http://dbpedia.org/property/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT count distinct ?dbresource where
  1. {?aWord wn20schema:lexicalForm "stadium".
  2. ?aWordSense wn20schema:word ?aWord .
  3. ?aSynsetRKB wn20schema:containsWordSense ?aWordSense .
  4. ?aSynsetRKB owl:sameAs ?aSynsetWN.
  5. ?dbresource dbprop:wordnet_type ?aSynsetWN.}
```

SPARQL listing 5.5: Query1 graph pattern from WordNet to DBpedia

Therefore in the pattern shown in *Figure: 5-4 Extracting synonyms using WordNet RDF/OWL* we reach the synset and then we leave WordNet and continue our graph pattern through line 5 in *SPARQL listing 5.5: Query1 graph pattern from WordNet to DBpedia*. Line 5 expresses the bridge between

WordNet and DBpedia. The variable ?dbresource in line 5 collects all DBpedia instances selected starting from the keyword “stadium” and no other restrictions have been set at this point. To evaluate the fitting of the DBpedia resource to the present work the number and the type of all the instances represented by ?dbresource have been requested. Firstly, all DBpedia resources that are selected without restricting the query to the instances that are mapped with LinkedGeoData have been requested. There are 2.663 DBpedia types corresponding to the synset coming from “stadium” ranging from <http://www.w3.org/2002/07/owl#Thing> to single buildings. Adding the lines 6 to 8 to the preceding query, we cross the bridge and enter into the LinkedGeoData ontology via the sameAs mappings:

```
PREFIX wn20schema: <http://www.w3.org/2006/03/wn/wn20/schema/>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX dbprop: <http://dbpedia.org/property/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX lgdvocab: <http://linkedgedata.org/vocabulary#>
SELECT distinct ?name where
  1. {?aWord wn20schema:lexicalForm "stadium".
  2. ?aWordSense wn20schema:word ?aWord .
  3. ?aSynsetRKB wn20schema:containsWordSense ?aWordSense .
  4. ?aSynsetRKB owl:sameAs ?aSynsetWN.
  5. ?dbresource dbprop:wordnet_type ?aSynsetWN.
  6. ?dbresource owl:sameAs ?geomapped.
  7. ?geomapped a ?type.
  8. ?geo lgdvocab:name ?name}
```

SPARQL listing 5.6 Query1 graph pattern - part 1 from WordNet to mapped instances

Line 6 is the bridge between DBpedia and LinkedGeoData. Line 7 identifies the type of data in LinkedGeoData that pertains to the term initially queried to WordNet (stadium) in LinkedGeoData. Line 8 identifies the names of the LGD classes identified in line 7. Moreover, referring to the LGD graph, it directs the graph pattern towards the LGD graph. Since DBpedia owns mappings with several resources on the LOD, we had to constrain the flow of information towards the LGD graph. Starting from the stadium keyword, we have 302 mapped instances whose types are 42 including node and way, the OSM data primitives. All mapped types relate to stadiums and the other terms, which are obtained through the linguistic component. Following is the list of the 42 types of data identified through the reduced query. We also have OSM primitive node and way (highlighted red in the table).

Type
<a href="http://linkedgedata.org/vocabulary#node">http://linkedgedata.org/vocabulary#node</a>
<a href="http://linkedgedata.org/vocabulary#stadium">http://linkedgedata.org/vocabulary#stadium</a>
<a href="http://linkedgedata.org/vocabulary#way">http://linkedgedata.org/vocabulary#way</a>

<a href="http://linkedgeodata.org/vocabulary#hockey%3Bbasketball%3Blacrosse">http://linkedgeodata.org/vocabulary#hockey%3Bbasketball%3Blacrosse</a>
<a href="http://linkedgeodata.org/vocabulary#basketball">http://linkedgeodata.org/vocabulary#basketball</a>
<a href="http://linkedgeodata.org/vocabulary#baseball">http://linkedgeodata.org/vocabulary#baseball</a>
<a href="http://linkedgeodata.org/vocabulary#football">http://linkedgeodata.org/vocabulary#football</a>
<a href="http://linkedgeodata.org/vocabulary#american+football+%3B+athletics">http://linkedgeodata.org/vocabulary#american+football+%3B+athletics</a>
<a href="http://linkedgeodata.org/vocabulary#soccer">http://linkedgeodata.org/vocabulary#soccer</a>
<a href="http://linkedgeodata.org/vocabulary#building">http://linkedgeodata.org/vocabulary#building</a>
<a href="http://linkedgeodata.org/vocabulary#hockey">http://linkedgeodata.org/vocabulary#hockey</a>
<a href="http://linkedgeodata.org/vocabulary#basketball%3Bhockey">http://linkedgeodata.org/vocabulary#basketball%3Bhockey</a>
<a href="http://linkedgeodata.org/vocabulary#football%3Bsoccer">http://linkedgeodata.org/vocabulary#football%3Bsoccer</a>
<a href="http://linkedgeodata.org/vocabulary#american_football">http://linkedgeodata.org/vocabulary#american_football</a>
<a href="http://linkedgeodata.org/vocabulary#hockey%3Bbasketball">http://linkedgeodata.org/vocabulary#hockey%3Bbasketball</a>
<a href="http://linkedgeodata.org/vocabulary#public_building">http://linkedgeodata.org/vocabulary#public_building</a>
<a href="http://linkedgeodata.org/vocabulary#football%3Bsoccer%3Brugby+league">http://linkedgeodata.org/vocabulary#football%3Bsoccer%3Brugby+league</a>
<a href="http://linkedgeodata.org/vocabulary#attraction">http://linkedgeodata.org/vocabulary#attraction</a>
<a href="http://linkedgeodata.org/vocabulary#construction">http://linkedgeodata.org/vocabulary#construction</a>
<a href="http://linkedgeodata.org/vocabulary#multi">http://linkedgeodata.org/vocabulary#multi</a>
<a href="http://linkedgeodata.org/vocabulary#rugby">http://linkedgeodata.org/vocabulary#rugby</a>
<a href="http://linkedgeodata.org/vocabulary#athletics%3B+soccer">http://linkedgeodata.org/vocabulary#athletics%3B+soccer</a>
<a href="http://linkedgeodata.org/vocabulary#Baseball">http://linkedgeodata.org/vocabulary#Baseball</a>
<a href="http://linkedgeodata.org/vocabulary#basketball%3Bvolleyball%3Bgymnastics%3B">http://linkedgeodata.org/vocabulary#basketball%3Bvolleyball%3Bgymnastics%3B</a>
<a href="http://linkedgeodata.org/vocabulary#Arena">http://linkedgeodata.org/vocabulary#Arena</a>
<a href="http://linkedgeodata.org/vocabulary#aussie_rules">http://linkedgeodata.org/vocabulary#aussie_rules</a>
<a href="http://linkedgeodata.org/vocabulary#cricket">http://linkedgeodata.org/vocabulary#cricket</a>
<a href="http://linkedgeodata.org/vocabulary#American+football">http://linkedgeodata.org/vocabulary#American+football</a>
<a href="http://linkedgeodata.org/vocabulary#basketball%3B+hockey">http://linkedgeodata.org/vocabulary#basketball%3B+hockey</a>
<a href="http://linkedgeodata.org/vocabulary#athletics">http://linkedgeodata.org/vocabulary#athletics</a>
<a href="http://linkedgeodata.org/vocabulary#australian_football">http://linkedgeodata.org/vocabulary#australian_football</a>
<a href="http://linkedgeodata.org/vocabulary#recreation_ground">http://linkedgeodata.org/vocabulary#recreation_ground</a>
<a href="http://linkedgeodata.org/vocabulary#athletics%3Bsoccer%3Blacrosse">http://linkedgeodata.org/vocabulary#athletics%3Bsoccer%3Blacrosse</a>
<a href="http://linkedgeodata.org/vocabulary#sport">http://linkedgeodata.org/vocabulary#sport</a>
<a href="http://linkedgeodata.org/vocabulary#icehockey">http://linkedgeodata.org/vocabulary#icehockey</a>
<a href="http://linkedgeodata.org/vocabulary#sports">http://linkedgeodata.org/vocabulary#sports</a>
<a href="http://linkedgeodata.org/vocabulary#university">http://linkedgeodata.org/vocabulary#university</a>
<a href="http://linkedgeodata.org/vocabulary#pedestrian">http://linkedgeodata.org/vocabulary#pedestrian</a>
<a href="http://linkedgeodata.org/vocabulary#rugby%3B+athletics">http://linkedgeodata.org/vocabulary#rugby%3B+athletics</a>
<a href="http://linkedgeodata.org/vocabulary#rugby%3Bfootball">http://linkedgeodata.org/vocabulary#rugby%3Bfootball</a>

<a href="http://linkedgeo.org/vocabulary#cyclling">http://linkedgeo.org/vocabulary#cyclling</a>
<a href="http://linkedgeo.org/vocabulary#football+cricket">http://linkedgeo.org/vocabulary#football+cricket</a>

**Table: 5.6 LinkedGeoData types of data from mapped "stadium"**

Unfortunately, together with Arena, that is the only synonym of stadium, we also have more general classes, like building, and other classes that are not strictly related to our task like university, cycling, pedestrian etc. Those nine types of LGD data have been highlighted yellow in the table. This kind of data seriously hinders the reliability of the results if we want to expand the query with a double instantiation assertion. From line one to eight, the query runs quickly and answers easily. Due to the mapping between LGD and DBpedia, following a graph pattern we are forced to enter into LGD through instances. At this point, unfortunately, the pattern cannot provide more. It has been proven that to obtain all instances that share the type of class or the type of property using a double instantiation assertion may lead to the extraction of an enormous quantity of useless data related to the yellow highlighted rows in Table: 4.6. A full query has been designed adding the two following lines 9 and 10 in the SPARQL query.

9. ?geo a ?type.
10.FILTER (?type != <http://linkedgeo.org/vocabulary#way> && ?type != <http://linkedgeo.org/vocabulary#node>)}

**SPARQL listing 5.7: Query 1 The graph pattern final lines**

Line 9 together with line 7 might be the double instantiation to extract all instances of classes involved in the mapping. Line 10 lets us discard OSM primitives from query results. Without the expansion of line 7 and 9, 302 stadiums have been obtained. Those stadiums are the ones that own a mapping with DBpedia. If we run the full query adding the last two lines of query1 the process slows dramatically since in line 7 we identified 42 types of data that we had already reported and analyzed in Table: 5.6 LinkedGeoData types of data from mapped "stadium" that might be queried. If we run the full query, we obtain more than two billion triples! The expansion involving types of instances that are not related semantically to the sample keyword is not a trustworthy procedure. It has already been underlined in *Table: 5.6 LinkedGeoData types of data from mapped "stadium"* the presence of classes that together with the bottleneck constituted by the low number of classes mapped between LinkedGeoData and DBpedia made this procedure disastrous.

### 5.5.1 Analysis of the results of query1 graph pattern

Query1 has been built starting from a lexical form browsing through linked datasets without any restriction on classes involved. The graph pattern drove us from WordNet through DBpedia to LinkedGeoData. Passing through the different resources on the LOD, we missed a lot of possibilities since the mapping between DBpedia and WordNet works for only 35% of WordNet, and

a considerably lower percentage of around 7% can be found for the mapping between LinkedGeoData and DBpedia. Lean mappings have been a considerable bottleneck for the flow of information. Another restriction came from the missing mappings between stadiums that are represented in DBpedia, who do not have a mapping with LGD. The last reduction of reliability came with the fact that in LGD many stadiums have been tagged with tags of a higher level of abstraction so by expanding the query we might obtain many (e.g.) buildings that are not stadiums. The only constraint that has been established in the query is the sample keyword because no other restriction on classes has been imposed. The query through the graph pattern resulted untenable. There is not a useful “short” path that we can follow starting from WordNet and arriving in LinkedGeoData. This is due to the nature of mappings that both ontologies have with DBpedia and the nature of DBpedia and LGD ontology. The graph pattern methodology used to build queries failed dramatically even though it is the only one that really tested the potential of the semantic web. The mappings with DBpedia are not suitable for accessing LinkedGeoData ontology for two reasons. Firstly, the mappings are stated for a few instances of both ontologies, secondly those instances pertain to a narrow number of classes. Moreover, due to the poorly structured LinkedGeoData Tbox, instantiation also occurred for overly general classes and the attempt to expand mappings from instances to classes lead to complete failure.

## 5.6 Query 2 - String matching

Browsing from the linguistic resource to geodata through interconnected resources in the Linking Open Data gave totally unsatisfactory results. The pattern worked only for one keyword over 15 and it worked only for the instances that owned a mapping between DBpedia and LinkedGeoData. The way LGD has been made blocks the possibility of expanding the mapping from the mapped instances to all the instances of a LinkedGeoData class.

Therefore, a second kind of query was designed. The second query uses a string matching method that drives directly from the extraction of synonyms in WordNet to LGD data. The query2 string matching avoids reliance on mappings with third party resources and inner taxonomies coming from folksonomies and works considerably better than the query1 based on the graph pattern. A graphical representation is given in the *Figure: 5-5 Query2 - String matching blocks* following page. A keyword (the single rhombus on the top left) is processed through WordNet. The results of the processing are a series of words that are synonymous words of the given keyword (yellow rhombuses on the left) and obviously the keyword itself. Through string matching, (red lines from lower left to lower right) over names of properties and classes in LGD, we can identify all the classes of real world objects (blue circles on top right) that are linguistically related through synonymy to the requested keyword. In this case the matching is between words and class or property names. DBpedia is then bypassed.

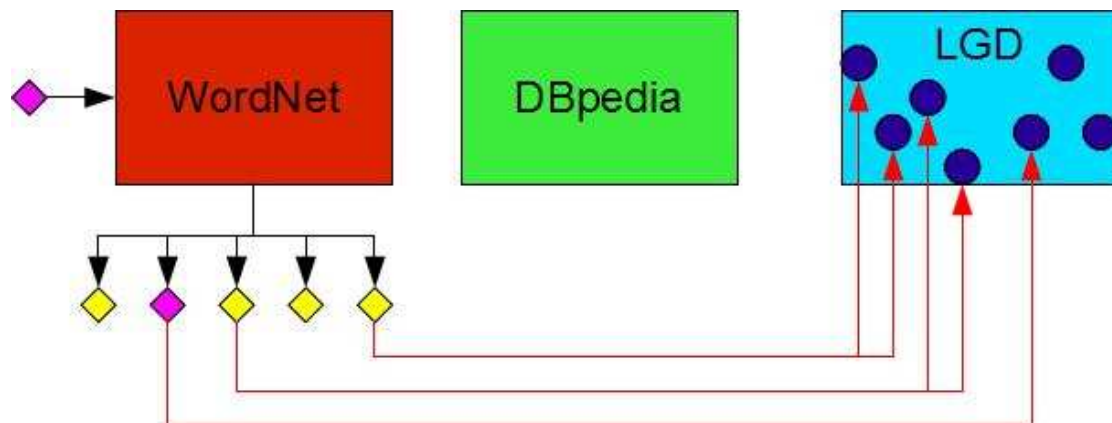


Figure: 5-5 Query2 - String matching blocks

### 5.6.1 Query construction

Bearing in mind that: The geodata coming from the query must come after a requester writes only a keyword to look for without any intermediate interaction, programming must be avoided since the semantic is embedded in the resources (or it might be). The possibility to use nested *SELECT* command that was introduced with SPARQL 1.1 (Harris & Seaborne, 2010) has been used to chain the two parts of the query to satisfy the above-mentioned constraints. This kind of query is very resource demanding because having to search over all the datasets involved without a pattern to follow, took hours to execute. Therefore fluxes of investigation have been directed to graphs using the *from <graph>* clause. This also has been favoured by the nested *SELECT* since both *SELECT* commands have been directed to a specific graph. The first running SPARQL query that was designed to satisfy the research aims is the following *SPARQL listing 5.8: Query2 String Matching*. Given a requested keyword, Query2 is able to return an answer with all the places in OSM that have been tagged with nouns semantically related to the requested keyword being synonymous with it:

```

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".}}}
```

SPARQL listing 5.8: Query2 String Matching



The query is composed of a main SELECT query that involves lines 2 to 3 and the nested linguistic query that involves lines 6 to 13. The linguistic query collects synonymous names in the variable ?lexforms. The variable ?lexforms and the filtering applied to LGD data in line 5 substitutes the path through DBpedia. In *query2 string match*, we only have the graph pattern inside the WordNet RDF/OWL resource as the only formalized ontology we are working with. The graph pattern between Lines 6 and 13 in *SPARQL listing 5.10: Querying LinkedGeoData through string match* has been explained in 5.4.3. Developing query2 string matching, the graph pattern explained in *SPARQL listing 5.4 Extracting synonymous words of bakery in WordNet RDF/OWL* has been followed entirely, extracting synonymous names and not stopping at the synset level as happened in query1. Going back to query2, from line 5 to line 1 the main query over LinkedGeoData is performed. The main query takes as its input the synonymous names from the nested lexical query. Line 5 and especially the variable ?lexforms is the red line between WordNet RDF/OWL and LinkedGeoData depicted in *Figure: 5-5 Query2 - String matching blocks*. Line 5 takes the values of the ?lexforms variable coming from the linguistic part. Together with line 4 asks LinkedGeoData ontology for all the elements that contain the lexical forms collected in ?lexforms variable. Line 4 stores in the ?geo variable all the instances/entities/features that in LGD ontology pertain to the classes identified through line 5 and we are thus going from class level to instance level. Line 3 is the *from <graph>* clause that identifies in the LOD the LGD resource. Line 2 identifies the variables whose values we want listed. In the following *Figure: 5-6 Query2 String matching, listing explanation* the role of the different lines of *SPARQL listing 5.8: Query2 String Matching* are represented.

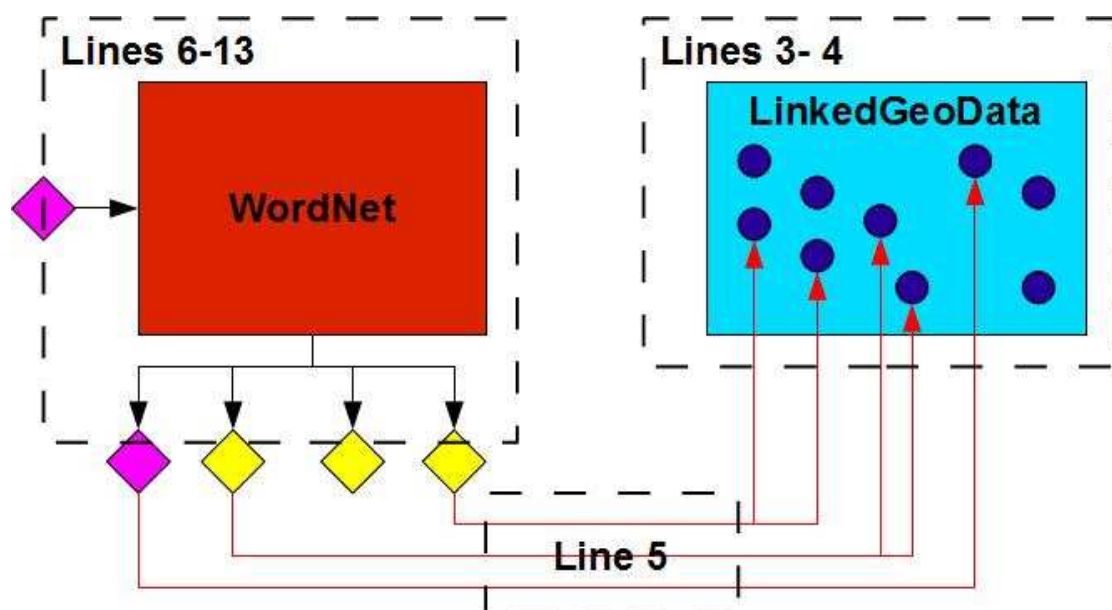


Figure: 5-6 Query2 String matching, listing explanation

### 5.6.2 Testing the query

To evaluate the coupling of the requested ontologies through the string matched query an evaluation based on the distinction of the two parts involved has been produced. The evaluation is composed of three parts.

The first part is composed of two phases: in the first phase (linguistic answers) all the sample keywords are submitted to the linguistic part to extract synonyms, in the second phase (geographic answers) all the starting sample keywords and the synonyms extracted in the first phase are searched in the LinkedGeoData graph. For the second part of the evaluation sample keywords are directly submitted to the full query2 *SPARQL listing 5.8: Query2 String Matching*. In the third part, the results of the two parts previously described are compared.

#### 5.6.2.1 Linguistic answers

In the first step, the sample keywords have been submitted to the linguistic query that we described in 5.4.3 and already partially used in the building of query1. Using the following query, the number of different synonyms of the queried term (lexical form) have been obtained.

```
PREFIX wn20schema:
<http://www.w3.org/2006/03/wn/wn20/schema/>
select count distinct ?lexforms
WHERE {?bWords wn20schema:lexicalForm ?lexforms.
?bWordSense wn20schema:word ?bWords .
?aSynset wn20schema:containsWordSense ?bWordSense .
?aSynset wn20schema:containsWordSense ?aWordSense .
?aWordSense wn20schema:word ?aWord .
?aWord wn20schema:lexicalForm "grocery".}
```

**SPARQL listing 5.9 Querying WordNet RDF/OWL for number of synonym of sample keywords**

Without the *count distinct* clause the synonymous terms of the given keyword can be obtained.

In Table: 5.7 Sample keywords synonyms, next page, the first column is for queried lexical form (keyword). When the queried lexical form is in WordNet RDF/OWL the term is in bold. In this case, if there are synonyms, they are reported in the second column. When no answer is given from WordNet RDF/OWL even when the requested lexical form is not in WordNet, the keyword is left normal (not bold).

Lexical form	Synonyms
<b>Grocery</b>	grocery store, market, foodstuff, food market
<b>Path</b>	way, way of life, route, course, track
<b>Parking</b>	
<b>Shelter</b>	Shelter, tax shelter, protection
footway <sup>70</sup>	
hiking <sup>71</sup>	
<b>Technology</b>	applied science, engineering, engineering science
<b>Frequency</b>	absolute frequency, relative frequency, frequency, oftenness
<b>Vending</b>	hawking, peddling, vendition
<b>Agricultural</b>	agrarian, farming
<b>Snowmobile</b>	
<b>Bakery</b>	bakeshop, bakehouse
<b>Network</b>	network web, electronic, meshing, meshwork, network mesh, network net
gritting <sup>69</sup>	
<b>Stadium</b>	sports stadium, bowl, arena

Table: 5.7 Sample keywords synonyms

### 5.6.2.2 Geographic answers

To find all results belonging to the synonyms of the original lexical form in LinkedGeoData the SPARQL endpoint has been queried. The bold column still underlines the role of the originating word. Moreover, queries have also been run for words that are not included in WordNet RDF/OWL.

The query is the following:

```
Select count distinct ?geo
Where {?geo a ?vocab
FILTER regex(str(?vocab), "grocery", "i")}
```

SPARQL listing 5.10: Querying LinkedGeoData through string match

The filter assures that we will obtain only terms in LGD that contain the keyword coming from the sample, grocery in the example above. The “i” flag assures that the match will be case insensitive. In *Table: 5.8 Synonyms of sample keywords and LinkedGeoData instances*, next page, information about a sample keyword and its synonyms are explained in two rows at a time. In the first row the sample keyword and its synonyms are reported, and in the second row the results of the above cited SPARQL query for the corresponding lexical forms are reported. The last column gives the total amount of entries found summing up all the single queries posted for the

<sup>70</sup> Still not in WordNet 3.0 in <http://www.oxfordadvancedlearnersdictionary.com/dictionary/footway> is classified like a formal British English noun since OSM uses tags in British English and WordNet is developed in the US it partially explains the absence of the term in WordNet.

<sup>71</sup> Inserted in WordNet 3.0

synonyms of a given sample keyword.

Lexical form	Synonymous of the given lexical form in WordNet RDF/OWL						Total
Entries in LGD	Entries of the synonymous in LGD						
<b>grocery</b> 192	foodstuff 0	food market 0	market 41.522	grocery store 0			41.714
<b>path</b> 95.596	way 3.800.508	way of life 0	Route 84	course 5.206	track 962.957		4.864.351
<b>parking</b> 191.325							191.325
<b>shelter</b> 4.395	tax shelter 0	protection 0					4.395
<b>footway</b> 669.487							669.487
<b>hiking</b> 1.418							1.418
<b>technology</b> 29	applied science 0	engineering 0	engineering science 0				29
<b>frequency</b> 0	absolute frequency 0	relative frequency 0	frequency 0	oftenness 0			0
<b>vending</b> 2.927	hawking 0	peddling 0	vendition 0				2.927
<b>agricultural</b> 7	agrarian 0	farming 1					8
<b>snowmobile</b> 4							4
<b>bakery</b> 7.239	bakeshop 1	bakehouse 0					7.240
<b>network</b> 2	network web 0	electronic 825	meshing 0	meshwork 0	network mesh 0	network net 0	827
<b>gritting</b> 0							0
<b>stadium</b> 5.244	sports stadium 0	bowl 1	Arena 17				5.262

Table: 5.8 Synonyms of sample keywords and LinkedGeoData instances

The sample keywords gritting and frequency have not been found in LinkedGeoData.

### 5.6.3 Query 2 String matching runtime

The LOD has been queried using the full query2 several times asking for different results concerning the sample. The query has already been reported in *SPARQL listing 5.8: Query2 String Matching*. The queried term after the Select clause in the second line has been substituted with (COUNT DISTINCT ?geo) to obtain the number of instances that the query selected for every sample keyword. Queries have been run using only all the sample keywords to see how many geographical entities (?geo) would be reached. Using the following as line two (select ?lexforms) to obtain how many synonyms (?lexforms) the queries gave the results described below.

The following *Table: 5.9 Query 2- string match results*. shows the results of the full query. The first column (lexical form) is the list of the queried sample keywords. The second column gives the synonyms of the given lexical forms that were found through WordNet RDF/OWL in the LGD knowledge base (querying select ?geo). The third column (Full query2) gives the number of entries that were found in the LGD database through the linguistic enrichment (querying select count distinct ?geo).

Lexical form	Synonymous	Full query2
Grocery	grocery store, market, foodstuff, food market	41.694
Path	way, way of life, route, course, track	6.058.107
Parking		191.325
Shelter	Shelter, tax shelter, protection	4.390
Footway		0
Hiking		0
Technology	applied science, engineering, engineering science	29
Frequency	absolute frequency, relative frequency, frequence, oftenness	0
Vending	hawking, peddling, vendition	2.926
Agricultural	agrarian, farming	7
Snowmobile		4
Bakery	bakeshop, bakehouse	7.238
Network	network web, electronic, meshing, meshwork, network mesh, network net	201
Gritting		0
Stadium	sports stadium, bowl, arena	5.262

**Table: 5.9 Query 2- string match results.**

The zero results for four sample keywords come from two different origins. Frequency and gritting were not found in the LGD knowledge base while footway, hiking and again gritting were not found in WordNet RDF/OWL.

#### 5.6.4 Analysis of the results of query 2 string matching

The *Table: 5.10 Evaluation of query2- string match*. The following page constitutes an evaluation framework to measure the performance of the query2 string matching against the results of a direct query without linguistic enrichment and against a sequence of the two queries that have been nested in query2. A couple of indicators have been added to support the evaluation of the performance of query2.

The first column (lexical form) is the list of the queried sample keywords.

The second column (Direct LGD) gives the number of entities that directly match the queried sample keyword in the LGD knowledge base without linguistic enrichment.

The third column (Full query2) gives the number of entries that have been found in the LGD database through the linguistic enrichment as in 5.6.3.

The fourth column (Direct LGD+) gives the number of entities that directly match the queried word and its synonym in the LGD knowledge base using the linguistic resource indirectly. It is the summing up of several queries posted directly to LinkedGeoData knowledge base as in 5.6.2.2.

The fifth column (Increase) gives the percentage of variation between the two queries that run once for every sample keyword. It has been evaluated as the percentage increase of results between a query searching directly the sample keyword on LinkedGeoData as in *SPARQL listing 5.10: Querying LinkedGeoData through string match* and the results of the query expansion using the full query2:

$$Increase = \frac{Query2 - DirectLGD}{100}$$

The sixth column (SWloss) gives a measure of missing results ignoring the link in the semantic web and querying directly LGD with the synonym words extracted previously using WordNet (either semantic or not).

It has been calculated according to the following:

$$SWloss = \frac{(DirectLGD+) - (Fullquery2)}{DirectLGD+}$$

As an example:

If we query “grocery” using the query2 we obtain 41.694 entries because the linguistic nested query added synonyms to the original lexical form “grocery”. Since querying LGD directly with the sample keyword we had only 192 “grocery” using the linguistic enrichment we increased the number of entries by 415,02%. If instead of going through the query2 we have the synonyms and we query LGD for all of them the divergence with the query2 is less than 0,00%.

Lexical form	Direct LGD	Full query2	Direct LGD+	Increase (%)	SWloss (%)
grocery	192	41.694	41.714	415,02	0,00
path	95.596	6.058.107	4.864.351	56.925,11	-0,25
parking	191.325	191.325	191.325	0,00	0,00
shelter	4.395	4.390	4.395	-0,05	0,00
footway	669.487	0	669.487	-6.694,87	1,00
hiking	1.418	0	1.418	-0,014	1,00
technology	29	29	29	0	0,00
frequency	0	0	0	-	-
vending	2.927	2.926	2.927	-0,01	0,00
agricultural	7	7	8	0,00	0,13
snowmobile	4	4	4	0,00	0,00
bakery	7.239	7.238	7240	-0,00	0,00
network	2	201	827	1,99	0,76
gritting	0	0	0	-	-
stadium	5.244	5.262	5262	0,0002	0,00

Table: 5.10 Evaluation of query2- string match.

Looking at the last table some considerations arose:

- *Mixture*: We mixed different kinds of entities since the sample is composed of classes and datatype properties as shown in *Table: 5.3 Sample keywords: types, subclasses superclasses*
- *Transportation network*: WordNet is a general-purpose linguistic resource; therefore, specialized terms for transportation networks entail slight differences from the common people's point of view. From the linguistic point of view, terms that in LGD and OSM<sup>72</sup> are strictly distinct in WordNet are considered synonymous, thus leading to an explosion of results querying entities like "path" that in WordNet is classified as a synonym of way.
- *Wordnet development*: We queried an "old" (according to obsolescence in IT terms) version of WordNet. Version 2.0 of Princeton WordNet was issued in 2003, possibly a more recent version entailing more contributions can lead to a more comprehensive vocabulary (WordNet 3.0 already includes hiking and gritting that were not in WordNet 2.0) or to a more finely grained distinction between terms describing transportation networks.
- *Shops*: When we queried shops (grocery, bakery, technology), we had good results without losing terms. We had a considerable increase of results for grocery since in WordNet has market amongst its synonyms.
- *Semantic web coupling*: Results that have been collected in the last column support the coupling of the two semantic resources through the nested query. Some instances have been lost but the loss almost never reaches one point in percentage.

<sup>72</sup> OSM has still in its name the purpose that driven it initial development, create an open street database to implement in an open navigation system. The general purpose map with areas and all details of maps came afterwards.

We can then say that, as pointed out previously, **the purpose of an ontology has to be considered in every kind of application we intend to use it in, the coupling of WordNet and LGD is suitable in the fields where the layperson’s language meets the layperson’s representation of space.**

## 5.7 Overall reflections on chapter five results

The query was constructed to make data collected through OpenStreetMap more usable, overcoming the low thematic accuracy. Below is a comparison of the results achieved in this chapter: a contextualization of the results both at development and at application level.

### 5.7.1 Query results comparison

An overall evaluation of the queries results can be derived looking at the following table:

Sample keywords	Synonyms	Used Synonyms	Direct LGD	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

**Table: 5.11 Overall evaluation of queries results**

In the second column we have the number of the synonyms including the given keyword. In the third column we have the synonyms of the given keyword used in LinkedGeoData. In the fourth column we have the number of entities that directly match the queried sample keyword in the LGD knowledge base without linguistic enrichment. In the fifth and seventh columns we have the results of the two developed queries. Thematic inaccuracy amongst the selected sample keywords was not so problematic with only 50% of the samples affected in some way. From a data consumer perspective, the enrichment is noticeable. Since a data consumer is potentially not an OSM



mapper and therefore might not know the content of standardization efforts inside OSM community achieved and documented through the map features list. Moreover only in two cases (grocery and path keywords) was a meaningful enrichment of the query found. Comparing the results shown in *Table: 5.8 Synonyms of sample keywords and LinkedGeoData instances*, we should note that the results of the enriched query have to be considered carefully according to the remarks of the previous section. The factors that foster the usability of crowdsourced data as identified by (Bishr & Kuhn, 2007) are part of the OpenStreetMap project as has already been underlined (last sentences of 3.4). Those factors consist in the availability of meta information by whom and when edited, a strong sense of community and the existence of a network of trusted users. Those factors and the fact that only standardized tags are rendered through rendering engines, enforced an inner standardization effort. The performance of the two designed queries is completely different. Query1 through graph pattern seems a failure. It worked only for one sample keyword and even in this case, it was not possible to expand the results to all the instances of the class stadium. It happened because tags pertaining to more abstract classes, like construction, building and similar, as listed in *Table: 5.6 LinkedGeoData types of data from mapped "stadium"*, own directly the instantiation of stadium instances hindering the attempt to expand the queries using a double instantiation. Query2, using the string matching overcame the limitation of mappings and multilayered instantiations and worked properly. As semantic resources are linked and structured now, query2 could be the core component of web applications that successfully manipulates OSM data and helps data consumers to find their desired locations.

### 5.7.2 The core component long lasting development

A core component has been developed. It has been developed relying on sparsely and differently generated applications stemming from different generations of the web and technologies involved as outlined in *Web 3.0 or semantic web* and in *Figure: 1-1 The evolution of web and related technologies*. From (Berners-Lee, Hendler, 2001).

The present thesis takes advantage of well-structured applications developed in the 90s and properly translated in semantic web resources in the first decade of the 21<sup>st</sup> century like WordNet. The core component combines WordNet with Web 2.0 applications started in 2000 (Wikipedia) to 2004 (OpenStreetMap). The core component developed in the web 3.0 environment takes advantage of recent semantic translations of the above cited resources Wordnet RDF/OWL, dated 2006, DBpedia, dated 2007, and LinkedGeoData, dated 2009.

The parts that have been coupled and their evolution in the following *Figure: 5-7 Evolution of resources towards the core component*. Where for every resource in boxes you can read the year of the first issue and the year of the semantic translation.

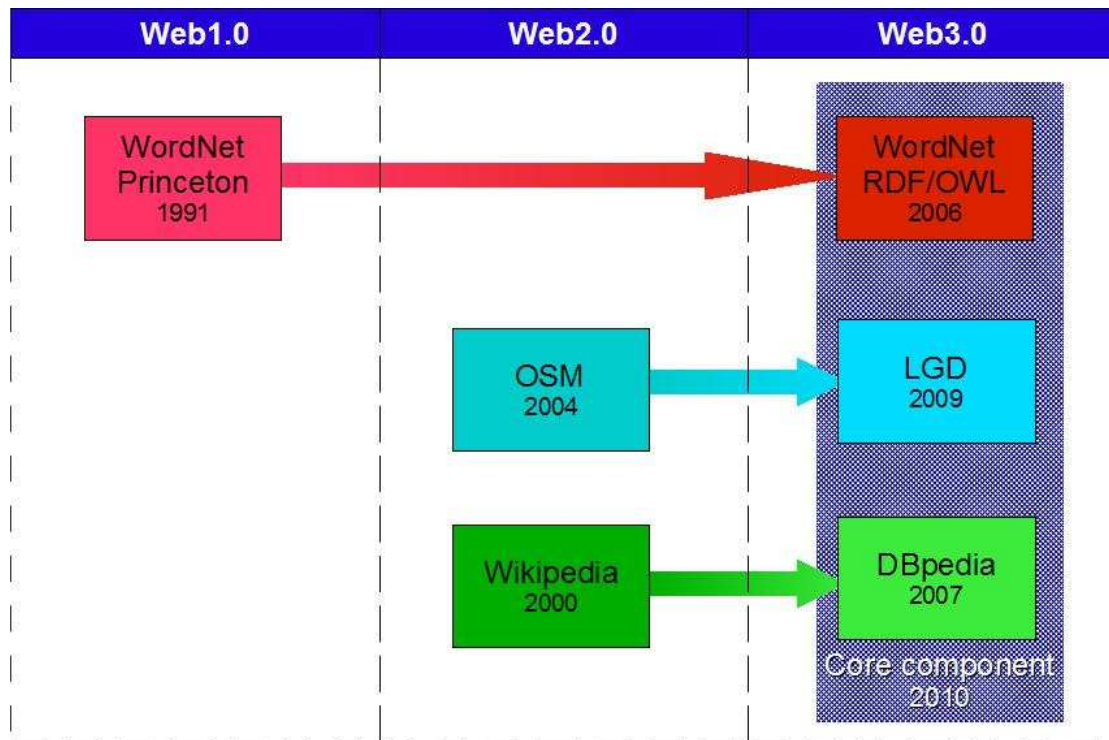


Figure: 5-7 Evolution of resources towards the core component.

### 5.7.3 Using the core component for an application

The core component favours the job of an application developer willing to use data collected by OpenStreetMap community for an application for mobile devices. The application might help users to find map locations they are looking for. The requester might have the possibility to insert a keyword on his mobile device and return a map with requested locations.

In the *Figure: 5-8 Application envisaged*, on the next page, the role of the core component in the above mentioned situation is depicted.

Two actors are shown in the figure below. The requester that holds the mobile device and the application developer that integrates the core component developed in this thesis with the other parts to obtain a running application represented in the following *Figure: 5-8* with the dashed line. The application developer needs to integrate the core component with a GUI to let the requester insert the keyword in the core component. Another element that might be developed is the rendering part. The application developer might integrate a system to convert the results of the query into mapped elements. The elements coming from the query could be either points, lines or areas expressed in RDF. The final part of the workflow might convert RDF geodata into map elements (red triangles) overlaid on a background map.

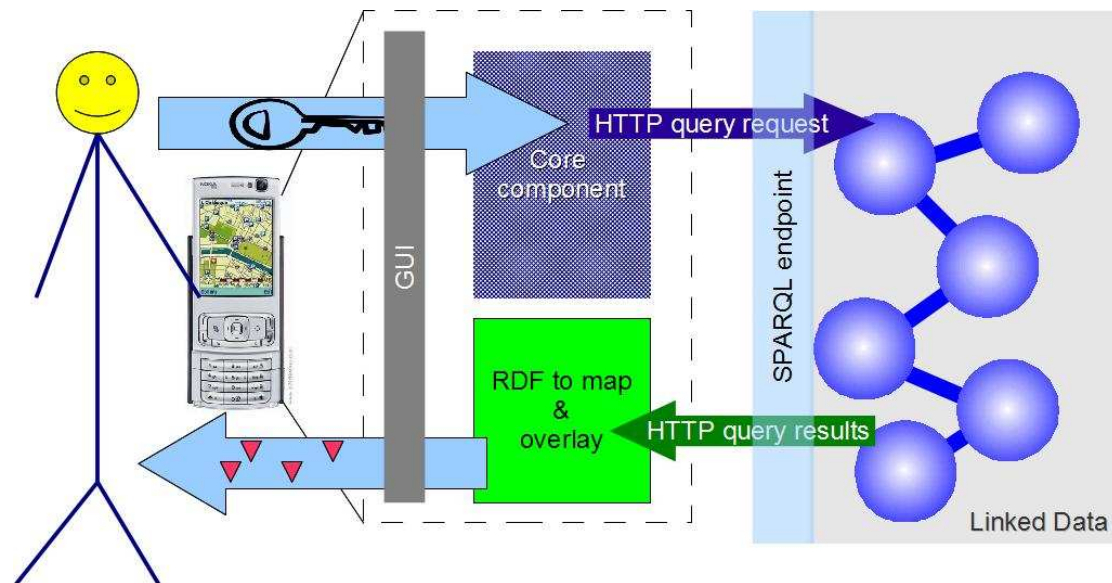


Figure: 5-8 Application envisaged

The workflow might look as follows. The requested keyword through the GUI is inserted in the core component. The core component through the LOD SPARQL endpoint sends a query request (query2 as SW resources are now) to the LOD. The query results coming from the LOD entails coordinates of points and areas in RDF format. Query results are then translated from RDF to map elements and then overlaid with a background map. Another possible application of the core component is the integration into the SPIRIT search engine. The semantic search engine that has been cited in 4.1 is intended to find geographical information from websites. In SPIRIT a spatial query expansion procedure is designed. The envisaged query expansion of SPIRIT is focused only on spatial containment hierarchies or alternative place names; there is not a module for a stricter linguistic expansion. This kind of query expansion might be provided by the core component developed here.

## 6 Chapter six: Conclusions and recommendations.

### Introduction

This thesis focused on the improvement of Volunteered Geographical Information. To achieve the research objective described below a “core component” has been developed. The core component consists of a query expansion obtained through a semantic query over semantic resources published in the web of data. The chapter is broken down thus: a deeper understanding of the research objectives in section 6.1; in section 6.3 some reflections are presented; in section 6.4 together with the answers to research questions an explanation of the main results achieved; Guidelines to choose between the two procedures to query building procedures in section 6.2, while Recommendations and Further developments are stipulated in section 6.5.

### 6.1 The main objective of the thesis

The main objective of the thesis is expressed by the following statement:

***Enhancement of the usability of a VGI dataset achieved overcoming the low thematic accuracy using semantic technologies in the environment of the semantic web.***

The initial general objective has always been the **Enhancement of a VGI dataset**. During the development of the thesis greater insights have determined certain refinements of the research objectives and questions. In fact, the other keywords involved in the research objective came following a waterfall procedure after analysis and comparisons between alternatives. The chosen VGI dataset was **OpenStreetMap** as explained in section 3.1 *OpenStreetMap and VGI*. The chosen enhancement was **usability** achieved via improved **thematic accuracy** as explained in section 1.5 *VGI data quality analysis*. This is followed by a discussion on the modelling of thematic accuracy issues as per section 1.6 *Modelling the problem: semantic heterogeneity in multiple databases* and a review on solution strategies in section 1.7 *Naming conflict: solving approaches* **semantic technologies** were chosen. The environment where semantic resources used for the development of the thesis are stored and linked is the **semantic web** presented in section 1.8. A pivotal role in the creation of the **semantic web** is the LinkingOpenData (LOD) initiative as in section 2.5. Moreover, the publication of LinkedGeoData (LGD) in the LOD, as in section 4.4, allowed us to present a more operational profile to the thesis. The LOD is where all the resources used in the development of the thesis are stored and linked.

## 6.2 Reflection

The thesis outlined the gap existing between two communities. The first community regards volunteer mappers. Volunteer mappers are boosted by the visibility of their edits on maps (Elwood, 2008b). In OpenStreetMap rendered information is only about a wide but still restricted number of tags listed in the “map features list” and this supports thematic accuracy. Mappers therefore are generally neither interested in the development of background information on rendered objects nor in the relation between tags either spatial (topological relations) or semantic.

The second community encountered regards the information specialists and geographers or geoinformation specialists. The second community is more focused on information ordering and meanings in a deeper way. The second community models using different abstraction levels and taxonomies to develop very complex knowledge bases ranging from general terms pertaining to upper level ontologies to application level ontologies (Guarino, 1998).

This project tried to explore the integration of the outcomes coming from both environments. LinkedGeoData as described in section 4.4 although developed by IT specialists was focused mainly on the translation and integration of OSM database in the semantic web. LGD inherited almost all the characteristics of the VGI community that developed OSM project. Therefore LGD has missing relationships between objects as you can notice in *Table: 4.2 Geographic information ontology elements and LinkedGeoData* and a weak taxonomy was developed.

LinkedGeoData developers created the taxonomy interpreting the OSM wiki pages and translating tags in class properties and literals as explained in section 4.4.2. Therefore, LGD taxonomy resulted in being poorly structured. In LinkedGeoData taxonomy you have few abstraction levels as reported in section 4.4.5. Moreover the class hierarchy has some strange relation between classes (e.g. the aeroway class and the highway class are subclasses of themselves as you can see in *Figure: 4-2 LinkedGeoData upper level hierarchy*) and some instantiation resulted as being wrong (e.g. some stadium classified in subclasses of shop). LinkedGeoData class hierarchy needs to be reconsidered in further research.

WordNet RDF/OWL developed by IT specialists pertaining to the second community structures information perfectly where the synonymy relation is only one of the possible amongst terms and concepts as explained in section 2.6.1 *WordNet data model*.

Linking these two efforts through the semantic web constituted the bridge trying to fill the gap between poorly structured VGI created data and the semantic web.

Other efforts attempted to connect the two perspectives.

The above cited attempts (in sections 4.3 and 4.5) developed but also

abandoned inside the community tried to use semantic technologies to give more semantic to the project.

Some projects developed in the second environment, apart from LinkedGeoData, are tracing the direction and developing the tools to overcome the limitations of LinkedGeoData both from the folksonomic point of view as reported in 2.7 and from the geographic information one as reported in 4.1.

## **6.3 Main results of the thesis**

### **6.3.1 Main conclusions**

The thesis, while targeting the improvement of the usability of OSM database, demonstrated the potential of the semantic web in integrating geographical information and other forms of knowledge. Semantically expressed knowledge can derive from a wide range of participants: from knowledge engineers and philosophers involved in higher ontology development to the IT enthusiast performing mapping operations using his GPS device.

This thesis has demonstrated that the integration of used information sources in the semantic web has been limited by many factors, above all, the minimal design of published geodata used for the development of the thesis. Secondly, the weak pre-processing of folksonomic geodata left aside semantic information embedded in databases and applications. Thirdly, the lean mappings between ontologies constitute a bottleneck for semantic web navigation as evaluated in section 4.4.4. The last limiting factor consists of clumsy taxonomies pertaining to ontologies coming from folksonomies (sections 2.5.2 and 4.4.5). All of these factors forced the development of two different approaches for the development of the semantic query: the initial, more semantic, graph pattern approach (sections 5.4 and 5.5) which gave unsatisfactory results (as outlined in section 5.7.1); and the second, string matching approach developed in section 5.6. The second approach gave a less semantic query but was able to support the desired improvement of VGI collected data usability and to evaluate thematic accuracy of the selected dataset.

We are therefore on one hand limited by the weak mappings and on the other hand we are limited by the burden constituted by the resources required to obtain results. Until now, there is no ideal solution for resource coupling.

The construction of the queries was the most difficult task. It required deep insight into the inner structure and links between the ontologies involved. The most cumbersome part was the analysis and evaluation of semantic resources, the study of the ontologies and the links stated in the LOD.

### 6.3.2 Answering the research questions

The sub-objectives of the research and the research questions.

The main objective of the thesis was the enhancement of the usability of volunteered geographic information with the main question therefore being:

Has a way been found to improve the usability of volunteered geographic information?

Two kinds of bridges between geodata and linguistic resources targeting usability were established in chapter five. One more semantic bridge has been created tracing a path along semantically related information in section 5.5 . The query underperformed due to the above cited bottleneck. A less purely semantic but more performing bridge has been created in section 5.6. Both bridges were compared and evaluated in section 5.7. In the second case, the improvement of usability was targeted. Moreover, we have to underline the fact that thematic accuracy and usability of a VGI dataset are not only a technical outcome. Thematic accuracy and usability of crowdsourced data in general are also dependant on some characteristics relating more to the originating project governance and tools. In OSM, for instance, we found considerable thematic accuracy amongst the sample keywords selected to test the queries.

The first sub objective:

#### 1. *Identify and characterize a VGI initiative*

To achieve the outlined sub-objective the following research questions were posed:

##### 1.1. Why OpenStreetMap?

As described in section 3.1 OpenStreetMap has three winning characteristics. Firstly, it is the most popular and developed geoweb2.0 application. Secondly, data are available to download. Thirdly, OpenStreetMap does not pose any limit on the reusability of data according to the licenses CC-BY-SA and ODBL.

##### 1.2. Is there any abstraction level or hierarchical structure in the OSM database?

Yes, there is a very simple abstraction level: tags and geographic primitives are divided. The tables in the database as shown in section 3.5 are based on the primitives, nodes and ways, tags are inserted in nodes and way tables owning a specific table. No other hierarchy is embedded in the data structure. There is not an implicit taxonomy amongst tags. The only type of hierarchical structure amongst tags is found in the wiki pages.

The second sub objective:

#### 2. *Identify an enhancement for the VGI initiative*

To achieve the outlined sub-objective the following research question was posed:

### 2.1. What kind of improvement can be achieved with the present thesis?

Valuable works were conducted analyzing OSM data measuring the quality of geodata according to quality standards. In this context, thematic accuracy and thus usability remained an uncovered issue as underlined in section 1.5.

Usability was intended in a twofold sense. At first it was related to the thematic accuracy of OSM and LGD. Usable in terms of consistent tagging, avoiding the “naming conflict” that arises when two editors (volunteered geographers) use synonymic terms to tag the same object. In a broader perspective, out of the borders of the OSM community, the database is more usable because, as the core component has been designed, the generic requester of geodata can ignore the standardization effort inside the OSM community. Requests passing through the linguistic enrichment are no longer related to more or less popular or standardized tags used in OpenStreetMap. Therefore, a generic requester can query using a keyword never used in OSM and still obtain positive results.

The third sub objective:

### 3. *Explore the relation between the chosen VGI initiative and the semantic technologies*

To achieve the outlined sub-objective the following research questions were posed:

#### 3.1. Why semantic technologies?

The thematic inaccuracy that arose during the merging of heterogeneous databases in section 1.6 was assumed as the best approximation of the thematic inaccuracy that arises in the development of user collected geodata. To solve the so called “naming conflict” that regards the use of synonymous tags for the same object, the semantic description of data has a pivotal role in the integration of heterogeneous databases.

#### 3.2. Has the broad OSM community attempted semantic developments of the project?

In section 4.2 five attempts have been shown. Four of them, developed using semantic technologies, were analyzed; another initiative is trying to give more semantic without taking advantage of semantic technologies. Amongst those initiatives only LinkedGeoData is running. The others are still at a development stage and seem to have been abandoned.

#### 3.3. How to deal with ontologies coming from ongoing web 2.0 application?

Lots of research projects are focusing on the development of ontologies from



web 2.0 applications. Some of them were cited in section 2.4.5. Some recent effort is trying to integrate different approaches on the extraction of semantically hidden content in web 2.0 applications and to render this data available for integration in the semantic web. The present work coupled three ontologies, two of them (LGD and DBpedia) come from folksonomies while the third (WordNet RDF/OWL) is a formal ontology. The former come from a translation, a post processing activity, of existing web 2.0 data. They made a lot of data available in the LOD but, as we have seen, they often lack semantic structure of data. Class hierarchies as can be seen in *Figure: 2-4 Places in DBpedia ontology* and in *Figure: 4-2 LinkedGeoData upper level hierarchy* lack a correct layered distinction between general and more specific terms. This kind of information is not provided by taggers and so class hierarchies of those resources have to be derived by semantic resource developers. One of the success factors of web 2.0 applications is the freedom of tagging without any compulsory action. Standardizations are merely suggestions. It has been called users' empowerment. Ontologies from folksonomies have not been created through user actions. Users have not inserted information directly in the ontologies. WordNet RDF/OWL is the last attempt to translate semantically WordNet 2.0 resulting in a considerably reliable and well designed resource. Therefore, a user never needs to edit semantic data directly. Suggestions on how to create an ontology on an ongoing web 2.0 application is provided above in section 6.5.1.4.

The fourth sub objective:

#### 4. *Explore the potential of the semantic web for geoinformation*

To achieve the outlined sub-objective the following research questions were posed:

##### 4.1. Are the available resources in the Semantic web able to support the task of the present work?

The semantic web was the technological environment where the desired results were achieved applying not only semantic technologies but using semantically developed resources described in section 2.5. Semantic resources on the web were accessed and queried directly on the web avoiding the use of any local copies. Moreover semantic resources were used semantically and not only as a repository of data. The thesis demonstrates that the semantic web has great potential for geoinformation but the resources available need further development to become really effective for applications. The semantic resources LinkedGeoData DBpedia and WordNet were used. There are billions of data and ontologies, for every kind of purpose, that are published in the LOD at a daily rate. Unfortunately, the way ontologies from folksonomies are linked and structured are weakening the usability of the semantic web. A problem that arose during the development of the thesis was that of the shortening of the timeout for answers by the LOD SPARQL endpoint maintainers. Most of the queries that were submitted during this

work (mainly those used in section 5.6) are no longer possible. The LOD SPARQL endpoint maintainers dramatically reduced the allowed timeout for queries. It resulted in the impossibility running the more complex queries that once took hours to return answers.

#### 4.2. Which way can other information sources in the semantic web be coupled with geoinformation?

The geographic resource used in this work is directly linked only with DBpedia. In any case, the coupling of information in the LOD in this thesis was obtained in two different ways. The more semantic one that drove query1 as in section 5.4, relying on a path following semantics along the LOD, is smarter (if the semantic information is properly linked) and quicker. The second coupling was obtained avoiding semantic relations between ontologies and was used to develop query2, is less refined and more resource demanding.

The more semantic query (the query1) relies on the fact that all semantic resources in the Linking Open Data initiative, as explained in section 2.5, are linked with some other semantic resource. Data about the earth's surface is in several ontologies but the richest collection of geodata is LinkedGeoData. Unfortunately, LinkedGeoData was mapped only with DBpedia, moreover as reported in section 4.4.4 only data with lat/lon information was mapped. The strict rules LGD developers decided for mappings of data limited the mappings enormously. It constitutes a considerable bottleneck for geographical information sharing in the semantic web.

The query2 that relies on string matching between nouns was the only way to return working matches between ontologies regarding geodata and lexical information. Since string matching coupling is not constrained by semantic relations between resources, all kinds of information inserted in the LOD can be coupled through string matching with LGD or any other kind of geoinformation in the SW.

The problem is that the string matching methodology is too resource demanding, the query lasts a long time and often exceeds the timeout.

#### 4.3. Which way to choose if too many tags for the same object lead to enriched or confused data? (How to deal with tag inconsistencies?)

Tag inconsistency can be enrichment if the depth of information is coupled with the depth of the requesting linguistic resource as described in 5.6.4.

#### 4.4. In what terms does the present work remain valid if we apply it to a different geoweb3.0 ontology?

The two developed queries have very different structures. Query1 is deeply rooted in links and location of resources on the LOD. It cannot be easily reused if we are willing to link WordNet RDF/OWL and a different Geo ontology in the LOD. Query2, jumping from WordNet to the graph containing

geoinformation requires small amendments and can therefore be easily adapted to a different Geo ontology in the LOD.

## 6.4 Guidelines to choose between the two procedures to query building

To deploy a query, a specific analysis of available data and the exploration of patterns between data might be executed. As we have seen, there is not a “standardized” linking mechanism. Two datasets can be linked through a sameAs assertion as in section 5.4.1 or through a property (*wordnet\_type*) as in section 5.4.2, therefore to explore the feasibility of a query that embraces different semantic web resources we have to investigate both the data and the links between them. The first step consists in identifying the datasets that we want to employ in the combined query. The list of all published and linked datasets in the LOD is available at <http://ckan.net/group/lodcloud> and the graphic representation of the data and their links is available through the Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch at <http://lod-cloud.net/>.

The second step consisted in the selection of some keywords. It was very important since keywords were used like icebreakers.

The third step consisted in the detection of the links between resources. To find the path one attempt can be made using refinder (<http://refinder.semanticweb.org/>) that is an Interactive Relationship Discovery in RDF Datasets. Relfinder has not been used in the development of the present thesis; the direct analysis of datasets documentation was preferred. The links between semantic resources in the LOD were detected following URI by URI the possible paths the semantic flow can follow. Therefore, to collect for every URI involved, all information was tested step by step and some semantic search engines were used like the ones listed in Bizer, Heath, Berners-Lee, (2009) or more directly simple SPARQL queries were built when no other resource was available.

In this thesis, two different procedures were followed to make semantically published ontologies work together through a semantic query. The choice between the two methodologies relies on different factors.

- Kind of data representation in ontologies
- Kind of patterns between ontologies.

The first step might be to find a graph pattern between data. When a reasonable path between data in different ontologies can be found and followed we can rely more on semantic links between data (hoping ontologies are properly developed) and we can return results in reasonable time<sup>73</sup>. In the present work, it was proven that the existence of a semantic path does not mean that the queries based on the path will have meaningful results. Due to

---

<sup>73</sup> In the last three months it is the only working possibility since the runtime for query has been shortened.

the lack of links in quality and in quantity and due to the absence of some measure of the quality of links, as ascertained in some OAEI initiative, it may so happen that queries pass over linked data and return no results as in section 5.5.1.

When a reasonable path is not possible between resources, the query developer has to search for different links between data. A naming triple with a simple plain text object can be good to link data. Otherwise a part of a text or a name of a class can be used to drive FILTER functions to find resources. In *SPARQL listing 5.8: Query2 String Matching* the FILTER function was used to select the class of LinkedGeoData instances.

To reduce the runtime the query has to be directed to graphs, otherwise without the guidance of the pattern the FILTER search will analyze all published datasets.

## 6.5 Recommendations and Further developments

Following the recommendations for further development of ontologies published in the LinkingOpenData initiative that were used for the development of the thesis. Further developments will focus on the envisaged extensions for the core component.

Recommendations are focused on an internally enhanced development of resources and on a better integration with other semantic web resources.

### 6.5.1 Recommendations for LinkedGeoData development

LinkedGeoData was evaluated in section 0 and section 4.7. Some limits of LinkedGeoData come from the strict derivation from OpenStreetMap like the difficult attempt to create a taxonomy from the poorly structured OSM database where semantic is missing. LGD developers wanted to translate a GI database in triples and then map it to an existing SW resource but inherited most of the original OSM data shortcomings. In the paper presenting the LinkedGeoData project<sup>74</sup> no reference is about both elicited ontologies from folksonomies and GI ontologies, notwithstanding those two fields have an enormous amount of ongoing research. In an LGD presentation paper, amongst 11 references, only one refers to something geographic with no reference to research about elicited ontologies from folksonomies.

#### 6.5.1.1 Elicitation of hidden semantic content

OSM is a GI folksonomy, therefore LGD developers had to extract the hidden semantic content in OSM (embedded either in the database or in the wiki or in the renderers) and validate it. LGD developers tried somehow to provide more semantics to OSM data but as shown in section 4.4.2 they obtained some misleading results. The processing of the original project as indicated in section 2.7 and the validation through the application of a reasoner to ascertain the validity and the consistence of the ontology is suggested.

---

<sup>74</sup> (Auer, Lehman, Hellman, 2009)

### **6.5.1.2 From a semantic geodata repository to a Geographic Information ontology**

Due to the way LGD was developed, it remains a taxonomy missing several aspects that might be included in geographic information ontology as underlined in section 0.

As any ontology from folksonomy, LGD can be considered a domain or a task ontology that needs an alignment with upper ontologies. The recognition of the proper abstraction level and some upper level concept can improve the usability of GI data to overcome the limits of the strict derivation from OSM. Since LinkedGeoData ontology is missing in upper level abstraction and in semantic relations, it can be a better geo ontology if it is better linked to external resources already inserted in the LOD.

LGD is already mapped with DBpedia but as was shown in section 2.5.2 DBpedia itself has a weak taxonomy and cannot be considered a performing upper level ontology.

For this reason and for a better integration in the semantic web, an improved policy for alignment with other semantic resources on the web is suggested.

### **6.5.1.3 Enriching mappings with other SW resources**

The mapping at instance level that involves LinkedGeoData and DBpedia, due to the scarcity of resources that have coordinates in DBpedia, and due to the fact that a lot of objects in real life that are mapped in OSM do not warrant a page in Wikipedia, is omitting a lot of useful data. Moreover, we found in literature that only 35% of WordNet owns a mapping with DBpedia. Mapping classes instead of instances only can be a solution but using OWL Full will reduce the usability of ontologies for automated processing. Moreover, the creation of links with resources other than DBpedia<sup>75</sup> is strongly suggested. The way DBpedia is related to WordNet as seen in section 5.4.2. could be a good example for a link based not only through a sameAs assertion. Instances of DBpedia are like instantiated in WordNet synsets and can be seen as a link between application ontology and an upper level ontology. Creating similar links between LinkedGeoData and WordNet may be a very important development for geodata in the web of data. A taxonomy based on WordNet<sup>76</sup> is considerably better than the one LGD developers elicited with their assumptions from OSM map feature list and will allow the mapping of almost all the LGD dataset.

### **6.5.1.4 Ontologies from folksonomies**

Another way to collect (geo)data from users can derive from a mixed approach to design ontologies for folksonomies. A semantic backbone and a

---

<sup>75</sup> The list of open data in the LOD at  
<http://esw.w3.org/TaskForces/CommunityProjects/LinkingOpenData/DataSets>

<sup>76</sup> A good overview of a taxonomy based on WordNet 3.0 for the term grocery  
<http://semanticweb.cs.vu.nl/europeana/session/thesaurus?query=grocery&thesaurus=http%3A%2F%2Fpurl.org%2Fvocabularies%2Fprinceton%2Fwn30%2F&type=http%3A%2F%2Fwww.w3.org%2F2000%2F01%2Frdf-schema%23Resource#lv=closed>

GUI that can help editors in the insertion of information in a kind of *snap to grid* approach aligning unstructured data to structured data as cited in section 2.4.5, where the grid might be an upper level ontology with general terms. LinkedGeoData is an ontology derived from the OpenStreetMap folksonomy so its grid might be developed stemming from research in both geographical information science and semantic translations of folksonomies as reported in section 2.7.

### **6.5.2 Further developments - Extending the core component with existing published resources**

As outlined previously the development of the present thesis is just a starting point for the integration of spatial information in the SW. Lots of developments can be envisaged for the applications outlined. For instance, one can be the development of queries involving the geocoded flickr pictures that have already been RDF serialized<sup>77</sup>. The application then might entail not only location for requested places but could also provide pictures of surrounding areas.

Another possible integration in the queries could be with multilingual ontologies. Some LOD resources already have multilingual descriptions for instances (DBpedia as well). This way a tourist might look for a place or activity using his native language without needing any knowledge of English. Other interesting resources can be the linkedsensordata<sup>78</sup>. It comprises data on weather stations in the US with coordinates for every station. On the LOD are also available all tweeter posts in TWARQL<sup>79</sup> linking them to the geographical aspect can be very fruitful for real time production and consumption of information, since TWARQL publishes on the LOD a few minutes after the Twitter post was created.

### **6.5.3 Extending the core component in the future: Geo SPARQL**

An important part of the application outlined in section 5.7.7 could be the possibility to integrate queries using functions set to manage geodata. There is not a standardized vocabulary and semantic for spatial queries on the semantic web, and various initiatives are on the go.

The first initiative is a standardization joint effort of OGC, ONTOLOG (SOCoP<sup>80</sup>) and W3C is hosted in OGC and a Special Work Group (SWG) was set up to coordinate the effort to define a spatial extension for SPARQL called GeoSPARQL (<http://ontology.cim3.net/cgi-bin/wiki.pl?GeoSparql>). A first draft document was presented in a slide set to explain the guidelines, purposes, resources and opportunities of the GeoSPARQL evolution (Lopez, 2010). Moreover a SWG charter was developed.

The second initiative is the possibility to evoke SQL MM commands in SPARQL queries developed by DBMS vendors. In OpenLink's Virtuoso

---

<sup>77</sup> <http://www4.wiwiss.fu-berlin.de/flickrwrapp/>

<sup>78</sup> [http://wiki.knoesis.org/index.php/SSW\\_Datasets](http://wiki.knoesis.org/index.php/SSW_Datasets)

<sup>79</sup> <http://wiki.knoesis.org/index.php/Twarql>

<sup>80</sup> Spatial Ontology Community of Practice

database system, the system that is hosting the LOD, there are SQL MM functions that can be evoked through a *bif:* namespace. This initiative is already running therefore we might create SPARQL queries over the LOD using the SQL MM functions. From September 2010, there is a SPARQL endpoint for LinkedGeoData, the access point is at (<http://linkedgeodata.org/sparql/>). In the LinkedGeoData webpage introducing the dedicated SPARQL endpoint there is an example query where a couple of geographic functions for queries are used. In Table 6.1 on the following page the complete list of SQL MM commands already available in OpenLink's Virtuoso. Those functions provide the possibility to add spatial constraints to our queries to restrict the queried terms to a selected area, to calculate distance, to calculate intersections and to query if there is containment between two geometries. Coming back to the application envisaged, there is the possibility to restrict the query to an area where the requester is (both from GPS embedded in the mobile device or mobile network cells in the case of a mobile phone), or an area the requester is able to identify (for this purpose the integration with geonames ontology can be envisaged) in case the requester is planning a trip.

The commands that were implemented for RDF in OpenLink Virtuoso are listed in Table: 6.1 SQLMM functions in Openlink Virtuoso below.

SQL MM function	Description
st_point	Returns a point geometry. The x coordinate corresponds to longitude.
st_x	Retrieves the x coordinate of a geometry.
st_y	Retrieves the y coordinate of a geometry.
st_distance	Returns the shortest distance between two points.
st_srid	Returns the srid of a geometry.
st_setsrid	The geometry given as argument is modified to have the specified srid and the modified geometry is returned.
st_astext	Returns the well known text (WKT) representation of the geometry.
st_geomfromtext	Parses the string and returns the corresponding geometry.
st_intersects	Returns intersects between two geometries.
st_within	Returns true if all points of a given geometry g1 are in another geometry g2.

isgeometry	Returns 1 if the argument is a geometry.
geo_insert	Inserts a geometry from an R tree index.
geo_delete	Deletes a geometry from an R tree index.
DB.DBA.RDF_GEO_ADD	Translates a geometry into a RDF box
DB.DBA.RDF_GEO_FILL	Converts geo:lat and geo:long properties into geometries.

**Table: 6.1 SQLMM functions in Openlink Virtuoso**



## 7 References

- Lots of literature on VGI can be found at the Workshop on Volunteered Geographic Information, December 13-14, 2007  
Goodchild Michael F., Gupta R. ,2007, *Workshop on Volunteered Geographic Information*, December 13-14, 2007 Santa Barbara, USA ,  
<http://www.ncgia.ucsb.edu/projects/vgi/>
- Abdelmoty, A. I, P. D Smart, C. B Jones, G. Fu, and D. Finch. 2005. *A critical evaluation of ontology languages for geographic information retrieval on the Internet*. Journal of Visual Languages & Computing 16, no. 4: 331–358.
- Angeletou, S., M. Sabou, L. Specia, and E. Motta. 2007. *Bridging the gap between folksonomies and the semantic web: An experience report*. In *Workshop: Bridging the Gap between Semantic web and Web*. Vol. 2.
- Assem M., Gangemi A., Schreiber G.,2006, *RDF/OWL Representation of WordNet W3C Working Draft* 19 June 2006
- Ather, Aamer. 2009. *A Quality Analysis of OpenStreetMap Data*. M.Eng. Dissertation, Department of Civil, Environmental & Geomatic Engineering University College London.
- Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, C., Ives, Z. 2007. *DBpedia: A Nucleus for a Web of Open Data*. Proceedings of the 6th International Semantic Web Conference (ISWC2007).
- Auer, Sören Lehmann, Jens Hellmann, Sebastian. 2009,1 *LinkedGeoData – Adding a spatial Dimension to the Web of Data*, Submitted to ISWC 2009.
- Auer Sören, Dietzold Sebastian, Lehmann Jens Hellmann, Sebastian Aumueller, David. 2009, . *Triplify – Lightweight Linked Data Publication from Relational Databases*, In Proceedings of WWW 2009
- Batini, C., M. Lenzerini, and S. B. Navathe. 1986. *A comparative analysis of methodologies for database schema integration*. ACM Computing Surveys (CSUR) 18, n° 4: 323–364.
- Berners-Lee, T., J. Hendler, O. Lassila, and others. 2001. *The semantic web*. Scientific American 284, no. 5: 28–37.
- Bishr Mohamed, Kuhn Werner, 2007 *Geospatial Information Bottom-Up: A Matter of Trust and Semantics* in (Fabrikant Wachovitz, 2007) 365-387

Gianfranco GlioZZo Msc GIMA Master thesis  
Bridging the gap between user generated spatial content and the semantic web

Spriengler Verlag

Bishr M., and L. Mantelas. 2008. *A trust and reputation model for filtering and classifying knowledge about urban growth*. *GeoJournal* 72, no. 3: 229–237.

Bizer Christian, Heath Tom, Berners-Lee Tim, 2009, *Linked Data - The Story So Far* International Journal on Semantic Web & Information Systems, Vol. 5, Issue 3 Information Resources Management Association ITJ5383

Bizer Christian; Lehmann Jens; Kobilarova Georgi; Auer, Sören; Becker, Christian; Cyganiak, Richard; Hellmann, Sebastian, 2009. *DBpedia - A Crystallization Point for the Web of Data*. *Web semantics*, Volume: 7, Issue: 3 (September 2009), pp: 154-165

Bray T., J. Paoli, C. M Sperberg-McQueen, E. Maler, and F. Yergeau. 2000. *Extensible markup language (XML) 1.0. W3C recommendation 6*.

Brickley D., R. V Guha, and B. McBride. 2004. *RDF vocabulary description language 1.0: RDF schema*. W3C recommendation 10: 27–08.

Buccella A., Cechich A. and Fillotrani P. 2009: *Ontology-driven geographic information integration: A survey of current approaches*, *Computers & Geosciences*, Special Issue on Geoscience Knowledge Representation in Cyberinfrastructure, 35, 4, pp. 710-723, 2009.

Buccella Agustina, Cechich Alejandra, Gendarmi Domenico, Lanubile Filippo, Semeraro Giovanni, Colagrossi Attilio, 2010: *GeoMergeP: Geographic Information Integration through Enriched Ontology Matching*, *New Generation Computing*, vol 28 issue 1 January 2010. Springer, Heidelberg

Buscaldi D., Rosso P., Sanchis E. 2006a, *WordNet-based Index Terms Expansion for Geographical Information Retrieval*. CLEF 2006 Working notes, Alicante 20-22 September, C.Peters Ed.

Buscaldi D., Rosso P., Sanchis E. 2006b: *Using the wordnet ontology in the geoclef geographical information retrieval task*. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 939–946. Springer, Heidelberg.

Castelli G., A. Rosi, M. Mamei, F. Zambonelli. 2007. *Ubiquitous Browsing of the World*. In (Scharl & Tochtermann, 2007) 67–78.

Chen W., Y. Cai, H. Leung, and Q. Li. 2010. *Generating ontologies with basic level. concepts from folksonomies*. *Procedia Computer Science* 1, n°. 1: 573–

Gianfranco Gliozzo Msc GIMA Master thesis  
Bridging the gap between user generated spatial content and the semantic web

581.

Clark K., G. Feigenbaum L. ,and Tortes,E, 2008 *SPARQL protocol for RDF*.  
W3C Recommendation 15 January 2008

Coote Andrew, and Les Rackham. 2008. *Neogeographic Data Quality is it an issue?* Paper presented at 2008 AGI conference. Stratford upon Avon, September 23, 2008.

Craglia M., . 2007. *Volunteered geographic information and spatial data infrastructures: When do parallel lines converge*. In Position paper for the VGI Specialist Meeting, Santa Barbara, 13–14.

Decker S. 2002. *Semantic Web and Databases: Relationships and some Open Problems*. Proceedings of the NSF-EU Workshop on Database and Information Systems: Research for Semantic Web and Enterprises, April 3 - 5, Amicalola Falls and State Park, Georgia

De Luca Ernesto, William Eul Martin and Nürnberger Andreas. , 2007. *Multilingual Query-Reformulation using an RDF-OWL EuroWordNet Representation*. In: Proceedings of the Workshop on Improving Web retrieval for non-English queries (iNEWS07). In conjunction with the SIGIR 2007 Konferenz, Amsterdam, 2007.

Devillers Rodolphe. and Jeansoulin, Robert., 2006. *Fundamentals of Spatial Data Quality* (Geographical Information Systems series) Wiley-ISTE, London.

Egenhofer, M. J. 2002. *Toward the semantic geospatial web*. In Proceedings of the 10th ACM international symposium on Advances in geographic information systems, 1–4.

Elwood S. 2008a. *Volunteered geographic information: key questions, concepts and methods to guide emerging research and practice*. GeoJournal 72, no. 3: 133–135.

Elwood, S. 2008b. *Volunteered geographic information: future research directions motivated by critical, participatory, and feminist GIS*. GeoJournal 72, no. 3: 173–183.

Fabrikant S. I., Wachowitz M. 2007: *The European Information Society* , Springer Verlag

Fallahi G. R, Frank A. U, Mesgari M. S, and Rajabifard A. 2008. *An ontological structure for semantic interoperability of GIS and environmental modeling*.

Gianfranco Gliozzo Msc GIMA Master thesis  
Bridging the gap between user generated spatial content and the semantic web

International Journal of Applied Earth Observations and Geoinformation 10,  
no. 3: 342–357.

Fellbaum Christiane ,1998 ed. *WordNet: An Electronic Lexical Database*.  
Cambridge, MA: MIT Press

Flanagin A. J, and Metzger. M. J, 2008. *The credibility of volunteered  
geographic information*. *GeoJournal* 72, no. 3: 137–148.

Fonseca, F., Davis C., and Câmara G. 2003. *Bridging ontologies and  
conceptual schemas in geographic information integration*. *Geoinformatica* 7,  
no. 4: 355–378.

Fonseca F., Egenhofer M., Davis C., and Câmara G. 2002. *Semantic  
granularity in ontology-driven geographic information systems*. *Annals of  
Mathematics and Artificial Intelligence* 36, no. 1: 121–151.

Fonseca F. T, Egenhofer M. J Agouris P., and G. Câmara. 2002. *Using  
ontologies for integrated geographic information systems*. *Transactions in GIS*  
6, no. 3: 231–257.

Gangemi A., and Mika P.. 2003. *Understanding the semantic web through  
descriptions and situations*. *Lecture Notes in Computer Science*: 689–706.

Gangemi, A. 2005. *Ontology design patterns for semantic web content*.  
*Lecture notes in computer science* 3729: 262.

Gangemi A., Catenacci C., Ciaramita M., Lehmann J., 2006 *Modelling  
Ontology Evaluation*, Y. Sure (ed.), *Proceedings of the Third European  
Semantic Web Conference*, Springer.

Gangemi A., Gliozzo A., Presutti V., Cardillo E., Daga E., Salvati A., and  
Troiani G.. 2007. *A Collaborative Semantic Web Layer to Enhance Legacy  
Systems*. *Lecture Notes in Computer Science* 4825: 764.

Gangemi A., Guarino N., Masolo C., Oltramari A., and Schneider L.. 2002.  
*Sweetening ontologies with DOLCE. Knowledge engineering and knowledge  
management: Ontologies and the semantic Web*: 223–233.

Gangemi A., Presutti V. 2009, *Ontology Design Patterns*, in Staab S. et al.  
(eds.): *Handbook of Ontologies* (2nd edition), Springer.

Guarino, N. 1998. *Formal Ontology in Information Systems: Proceedings of  
the 1st International Conference June 6-8, 1998, Trento, Italy*. Ios Press  
Amsterdam pp.3-15

Gianfranco Gliozzo Msc GIMA Master thesis  
Bridging the gap between user generated spatial content and the semantic web

Goodchild M. F., 1995. *Attribute accuracy*. In (Guptill & Morrison, 1995) Oxford, Elsevier, p 59–80

Goodchild M. F., 2007a. *Citizens as sensors: the world of volunteered geography*. *GeoJournal* 69, no. 4: 211–221.

Goodchild M. F. 2007b. *Citizens as voluntary sensors: spatial data infrastructure in the world of Web 2.0*. *International Journal of Spatial Data Infrastructures Research* 2: 24-32.

Goodchild M. F. 2008. *Commentary: whither VGI?* *GeoJournal* 72, no. 3: 239–244.

Gouveia, C., and Fonseca A., 2008. *New approaches to environmental monitoring: the use of ICT to explore volunteered geographic information*. *GeoJournal* 72, no. 3: 185–197.

Gruber T. R., 1993, *A translation approach to portable ontologies*. *Knowledge Acquisition*. Vol. 5, No. 2,

Gruber Tom, 2007a. *Ontology of folksonomy: A mash-up of apples and oranges*. *International Journal on Semantic Web and Information Systems* 3, n° 1: 1–11.

Gruber Tom, 2007b. *Collective knowledge systems: Where the social web meets the semantic web*. *Web Semantics: Science, Services and Agents on the World Wide Web* 6, n° 1: 4–13.

Gruber Tom, 2008, *Ontology*. *Entry in the Encyclopedia of Database Systems*, Ling Liu and M. Tamer Özsu (Eds.), Springer-Verlag.

Gruninger, M., Bodenreider O., Olken F., Obrst L., and Yim P.. 2008. *Ontology Summit 2007–Ontology, taxonomy, folksonomy: Understanding the distinctions*. *Applied Ontology* 3, n° 3: 191–200.

Guptill S.C. and Morrison J.L. (eds), 1995, *Elements of Spatial Data Quality*, Oxford, Elsevier.

Hakimpour, F., and Geppert A.. 2005. *Resolution of semantic heterogeneity in database schema integration using formal ontologies*. *Information Technology and Management* 6, n° 1: 97–122.

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

- Gianfranco Gliozzo Msc GIMA Master thesis  
Bridging the gap between user generated spatial content and the semantic web
- Haklay M., Singleton A. , Parker C. 2008. *Web mapping 2.0: the Neogeography of the Geoweb*. *Geography Compass* 2, no. 6: 2011–2039.
- Haklay M., and Weber P., 2008. *OpenStreetMap: user-generated street maps*. *IEEE Pervasive Computing* 7, no. 4: 12–18.
- Harding Jenny. 2005. *Vector Data Quality: A Data Provider's Perspective*. In (Devillers & Jeansoulin, 2005) Wiley-ISTE.
- Harding Jenny; Sharples Sarah; Haklay Muki; Burnett Gary; Dadashi Yasamin; Forrest David; Maguire Martin; Parker Christopher. J.; Ratcliff Liz., 2009, *Usable geographic information – what does it mean to users?* AGI GeoCommunity '09 Conference, Stratford-Upon-Avon;
- Harris Steve; Seaborne Andy eds. *SPARQL Query Language 1.1 W3C Working Draft* 26 January 2010
- Harvey F., Kuhn W., Pundt H., Bishr Y., and Riedemann C., 1999. *Semantic interoperability: A central issue for sharing geographic information*. *The Annals of Regional Science* 33, no. 2: 213–232.
- Howe J. 2006. *The Rise of Crowdsourcing* in *Wired*, Vol. 14, No. 6.
- Hull Richard, 1997. *Managing semantic heterogeneity in databases: a theoretical prospective*. In *Proceedings of the Sixteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems* (Tucson, Arizona, United States, May 11 - 15, 1997). PODS '97. ACM, New York, NY, 51-61.
- ISO/TC 211, 2002, *19113 Geographic information – Quality principles*, International Organization for Standardization (ISO), 2002.
- Kashyap Vipul, Bussler Christoph, Moran Matthew 2008 *The Semantic Web - Semantics for Data and Services on the Web*, Springer
- Kavouras, M., Kokla M. 2000. *Ontology-based fusion of geographic databases*. In *FIG Com3 Workshop and Annual Meeting*, Athens, Greece.
- Kim, H. L, Scerri S., Breslin J. G, Decker S., and Kim H. G., 2008. *The state of the art in tag ontologies: a semantic model for tagging and folksonomies*. In *Proceedings of the 2008 international Conference on Dublin Core and Metadata Applications*, 128–137.
- Kim H. L, Passant A., Breslin J. G, Scerri S., and Decker S. 2008. *Review and alignment of tag ontologies for semantically-linked data in collaborative tagging spaces*. In *The IEEE International Conference on Semantic Computing*, 315–322.

Gianfranco Gliozzo Msc GIMA Master thesis  
Bridging the gap between user generated spatial content and the semantic web

Klien E., and Probst. F. 2005, *Requirements for geospatial ontology engineering*. In 8th Conference on Geographic Information Science (AGILE 2005), Estoril, Portugal, 24–29.

Knerr T. 2006. *Tagging ontology-towards a common ontology for folksonomies*. Retrieved June 14: 2008.

Kuhn, W. 2005. *Geospatial semantics: Why, of what, and how?* Lecture Notes in Computer Science 3534: 1–24.

Lassila, O., and R. R Swick. S.d. *Resource Description Framework (RDF) Model and Syntax Specification*, W3C Recommendation 22 February 1999. W3C-World Wide Web Consortium,[Online] <http://www.w3.org/TR/REC-rdf-syntax>.

Lemmens, R. 2006. *Semantic interoperability of distributed geo-services*. Nederlandse Commissie voor Geodesie.

Lemmens, R., R. de By, M. Gould, A. Wytzisk, C. Granell, and P. van Oosterom. 2007. *Enhancing Geo-Service Chaining through Deep Service Descriptions*. Transactions in GIS 11, no. 6: 849–871.

Lemmens, R., C. Granell, A. Wytzisk, R. de By, M. Gould, and P. van Oosterom. 2006. *Semantic and syntactic service descriptions at work in geo-service chaining*. In Proc. of the 9th AGILE Int. Conference on Geographic Information Science. Visegrád, Hungary.

Lenat, D. B. 1995. *CYC: A large-scale investment in knowledge infrastructure*. Communications of the ACM 38, n°: 11: 33–38.

Li, W. S, and C. Clifton. 1994. *Semantic integration in heterogeneous databases using neural networks*. In Proceedings of the International Conference on Very Large Data Bases, 1–1.

Li, W. S, C. Clifton, and S. Y Liu. 2000. *Database integration using neural networks: implementation and experiences*. Knowledge and Information Systems 2, n°: 1: 73–96.

Lieberman, Joshua editor. 2006. *Geospatial Semantic Web Interoperability Experiment Report*. Open Geospatial Consortium Inc., May 7.

Lohar, S., 2010. *Semantic Integration of Heterogeneous Databases in Multidatabase System*. Master thesis in Computer Science & Engineering Thapar University, Patiala

Gianfranco Gliozzo Msc GIMA Master thesis  
Bridging the gap between user generated spatial content and the semantic web

Lopez, Xavier, 2010 *GeoSPARQL - A geographic query language for RDF data – A proposal for an OGC Draft Candidate Standard*. Available at [http://www.ogcnetwork.net/system/files/Spatial\\_SPARQL\\_Lopez.pdf](http://www.ogcnetwork.net/system/files/Spatial_SPARQL_Lopez.pdf)

de Man, W. H.E, and W. H van den Toorn. 2002. *Culture and the adoption and use of GIS within organisations*. International Journal of Applied Earth Observation and Geoinformation 4, no. 1: 51–63.

Mark, D., B. Smith, and B. Tversky. 1999. *Ontology and geographic objects: An empirical study of cognitive categorization*. Spatial Information Theory. Cognitive and Computational Foundations of Geographic Information Science: 747–747.

Masolo, C., S. Borgo, A. Gangemi, N. Guarino, A. Oltramari, and L. Schneider. 2003. *DOLCE: a descriptive ontology for linguistic and cognitive engineering*. WonderWeb Project, Deliverable D 17.

Maué, P., and S. Schade. 2008. *Quality Of Geographic Information Patchworks*. In . University of Girona, Spain .

Meersman, R. 2001. Ontologies and databases: More than a fleeting resemblance. Rome OES/SEO Workshop 14-15 September.

Mika, P. 2007. *Ontologies are us: A unified model of social networks and semantics*. Web Semantics: Science, Services and Agents on the World Wide Web 5, n° 1: 5–15.

Miller, George A. ,1995. *WordNet: A Lexical Database for English*. Communications of the ACM Vol. 38, No. 11: 39-41

Naiman, C. F, and A. M Ouksel. 1995. *A classification of semantic conflicts in heterogeneous database systems*. Journal of Organizational Computing and Electronic Commerce 5, n° 2: 167–193.

Obrst Leo, 2006 *The Ontology Spectrum & Semantic Models* Presentation for OntologySummit2007: A Collection of Definitions - What do people mean when they use the term "Ontology" available at [http://ontolog.cim3.net/cgi-bin/wiki.pl?ConferenceCall\\_2006\\_01\\_12](http://ontolog.cim3.net/cgi-bin/wiki.pl?ConferenceCall_2006_01_12)

OGC and ISO, 2001. *Abstract Specification Topic 1: Feature Geometry*. Available at <http://www.opengeospatial.org/standards/as>

OGC. 2003. *GML Geography Markup Language (GML) 3.0 Implementation Specification*. OpenGIS Consortium.



Gianfranco Gliozzo Msc GIMA Master thesis  
Bridging the gap between user generated spatial content and the semantic web

OGC , 2009. *Abstract Specification Topic 5: Features*. Available at <http://www.opengeospatial.org/standards/as>

Percivall, G. ed., 2008. *OGC Reference Model*. Open Geospatial Consortium, OGC Reference number: OGC 08-062r4 Version: 2.0

Peuquet, Donna J. 1988. *Representations of Geographic Space: Toward a Conceptual Synthesis*. Annals of the Association of American Geographers 78, no. 3: 375-394. doi:Article.

Presutti, V., and A. Gangemi. 2008. *Content ontology design patterns as practical building blocks for web ontologies*. In Proceedings of the 27th International Conference on Conceptual Modeling (ER 2008), Berlin.

Prud'hommeaux, Eric ;Seaborne, Andy eds. 2008 *SPARQL Query Language for RDF* W3C Recommendation 15 January 2008

Reddy, M. P., B. E. Prasad, P. G. Reddy, and A. Gupta. 1994. *A methodology for integration of heterogeneous databases*. Knowledge and Data Engineering, IEEE Transactions on 6, n° 6: 920–933 .

Reeve, D., and J. Petch. 1999. *GIS Organizations and People: A Socio-technical Approach* (UK: International Ltd, Padstow).

Schade, S., and P. Maué. S.d., 2008, *Standardizing the Geospatial Semantic Web?*Position Paper Workshop "Semantic Web meets Geopatial Applications", held in conjunction with AGILE

Specia, L., and E. Motta. 2007. Integrating folksonomies with the semantic web. The semantic web: research and applications: 624–639.

Scharl, A., and K. Tochtermann. 2007. *The geospatial web: how geobrowsers, social software and the Web 2.0 are shaping the network society*. Springer Verlag.

Schuurman, N., and A. Leszczynski. 2006. *Ontology-based metadata*. Transactions in GIS 10, no. 5: 709–726.

Servigne, S., N. Lesage, and T. Libourel. 2005. *Quality Components, Standards, and Metadata*. In (Devillers & Jeansoulin, 2005) Wiley-ISTE.

Sheth, A., and R. Meersman. 2002. *Amicalola Report: Database and Information Systems Research Challenges and Opportunities in Semantic Web and Enterprises*. Proceedings of the NSF-EU Workshop on Database and Information Systems: Research for Semantic Web and Enterprises, April 3 - 5, Amicalola Falls and State Park, Georgia

Gianfranco GlioZZo Msc GIMA Master thesis  
Bridging the gap between user generated spatial content and the semantic web

Smith, B. and Welty. C., 2001. *Ontology: Towards a new synthesis*. Proceedings of the international conference on Formal Ontology in Information Systems - Volume 2001 Ogunquit, Maine, USA 3 - .9, ACM New York, NY, USA

Spyns, Peter Meersman, Robert Jarrar, Mustafa, 2002 *Data modelling versus ontology engineering* ACM SIGMOD Record. Vol. 31, no. 4, pp. 12-17. Dec. 2002

Stoter, J., W. Quak, P. van Oosterom, M. Meijers, R. Lemmens, and H. Uitermark. 2007. *Considerations for the design of a semantic data model for a multi-representation topographical database*. Lecture notes in information sciences. Berlin: CODATA: 53–71.

Stuckenschmidt, H., and F. Van Harmelen. 2004. *Information sharing on the semantic web*. Springer-Verlag New York Inc.

Tapscott, D., and A. D Williams. 2006. *Wikinomics: How mass collaboration changes everything*. Portfolio.

Tomai, E., and M. Kavouras. 2004. *From “onto-geonoesis” to “onto-genesis”*: *The design of geographic ontologies*. Geoinformatica 8, no. 3: 285–302.

Tulloch, D. 2008. *Is volunteered geographic information participation*. GeoJournal.

Turner, A. 2006. *Introduction to Neogeography*. O'Reilly.

Unwin, D. J. 2005. *Fiddling on a different planet?* Geoforum 36, no. 6: 681–684.

Uschold Mike, Grüninger Michael, 1996, *Ontologies: principles, methods, and applications* Knowledge Engineering Review, Vol. 11, No. 2. , pp. 93-155.

van Assem M., Gangemi A., Schreiber G., 2006, *Conversion of WordNet to a standard RDF/OWL representation*, in Proceedings of LREC2006, Genova.

Van Damme, C., M. Hepp, and K. Siorpaes. 2007. *Folksontology: An integrated approach for turning folksonomies into ontologies*. Bridging the Gap between Semantic Web and Web 2: 57–70.

W3C OWL Working Group, 2009, *OWL 2 Web Ontology Language Document Overview*, W3C Recommendation 27 October 2009 available at <http://www.w3.org/TR/owl2-overview/> , W3C

Gianfranco GlioZZo Msc GIMA Master thesis  
Bridging the gap between user generated spatial content and the semantic web

Zhao, S., and E. Chang. 2007. *From database to semantic web ontology: an overview*. In *On the Move to Meaningful Internet Systems 2007: OTM 2007 Workshops*, 1205–1214.

Website references:

Berners-Lee Tim, Hendler James, 2001, *Scientific publishing on the 'semantic web*, Nature 410, 1023-1024, free available at <http://www.nature.com/nature/debates/e-access/Articles/bernerslee.htm>

Bratt, Steve, 2006, Emerging Web Technologies to Watch, W3C available at [http://www.w3.org/2006/Talks/1023-sb-W3CTechSemWeb/#\(1\)](http://www.w3.org/2006/Talks/1023-sb-W3CTechSemWeb/#(1))

OAEI, 2009, Ontology Matching OM-2009 Papers from the ISWC Workshop available at (<http://oaei.ontologymatching.org/doc/>)

O'Reilly, T. 2007. What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software (30 September 2005). [online] available at <http://events.oreilly.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html> .[Accessed 27 November 2009].

Vander Wal, Thomas. 2007. Folksonomy :: vanderwal.net. [online] available at <http://vanderwal.net/folksonomy.html> .[Accessed 28 July 2009]

## 8 Appendix: LinkedGeoData type of data with a mapped instance with DBpedia

Mapped Classes
<a href="http://linkedgeo.org/vocabulary#node">http://linkedgeo.org/vocabulary#node</a>
<a href="http://linkedgeo.org/vocabulary#town">http://linkedgeo.org/vocabulary#town</a>
<a href="http://linkedgeo.org/vocabulary#lighthouse">http://linkedgeo.org/vocabulary#lighthouse</a>
<a href="http://linkedgeo.org/vocabulary#building">http://linkedgeo.org/vocabulary#building</a>
<a href="http://linkedgeo.org/vocabulary#water">http://linkedgeo.org/vocabulary#water</a>
<a href="http://linkedgeo.org/vocabulary#way">http://linkedgeo.org/vocabulary#way</a>
<a href="http://linkedgeo.org/vocabulary#university">http://linkedgeo.org/vocabulary#university</a>
<a href="http://linkedgeo.org/vocabulary#city">http://linkedgeo.org/vocabulary#city</a>
<a href="http://linkedgeo.org/vocabulary#village">http://linkedgeo.org/vocabulary#village</a>
<a href="http://linkedgeo.org/vocabulary#aerodrome">http://linkedgeo.org/vocabulary#aerodrome</a>
<a href="http://linkedgeo.org/vocabulary#stadium">http://linkedgeo.org/vocabulary#stadium</a>
<a href="http://linkedgeo.org/vocabulary#peak">http://linkedgeo.org/vocabulary#peak</a>
<a href="http://linkedgeo.org/vocabulary#station">http://linkedgeo.org/vocabulary#station</a>
<a href="http://linkedgeo.org/vocabulary#country">http://linkedgeo.org/vocabulary#country</a>
<a href="http://linkedgeo.org/vocabulary#suburb">http://linkedgeo.org/vocabulary#suburb</a>
<a href="http://linkedgeo.org/vocabulary#post_office">http://linkedgeo.org/vocabulary#post_office</a>
<a href="http://linkedgeo.org/vocabulary#River">http://linkedgeo.org/vocabulary#River</a>
<a href="http://linkedgeo.org/vocabulary#school">http://linkedgeo.org/vocabulary#school</a>
<a href="http://linkedgeo.org/vocabulary#weir">http://linkedgeo.org/vocabulary#weir</a>
<a href="http://linkedgeo.org/vocabulary#airport">http://linkedgeo.org/vocabulary#airport</a>
<a href="http://linkedgeo.org/vocabulary#island">http://linkedgeo.org/vocabulary#island</a>
<a href="http://linkedgeo.org/vocabulary#coastline">http://linkedgeo.org/vocabulary#coastline</a>
<a href="http://linkedgeo.org/vocabulary#soccer">http://linkedgeo.org/vocabulary#soccer</a>
<a href="http://linkedgeo.org/vocabulary#Stream">http://linkedgeo.org/vocabulary#Stream</a>
<a href="http://linkedgeo.org/vocabulary#administrative">http://linkedgeo.org/vocabulary#administrative</a>
<a href="http://linkedgeo.org/vocabulary#football">http://linkedgeo.org/vocabulary#football</a>
<a href="http://linkedgeo.org/vocabulary#Observatory">http://linkedgeo.org/vocabulary#Observatory</a>
<a href="http://linkedgeo.org/vocabulary#runway">http://linkedgeo.org/vocabulary#runway</a>
<a href="http://linkedgeo.org/vocabulary#riverbank">http://linkedgeo.org/vocabulary#riverbank</a>
<a href="http://linkedgeo.org/vocabulary#basketball">http://linkedgeo.org/vocabulary#basketball</a>
<a href="http://linkedgeo.org/vocabulary#baseball">http://linkedgeo.org/vocabulary#baseball</a>
<a href="http://linkedgeo.org/vocabulary#american_football">http://linkedgeo.org/vocabulary#american_football</a>

<a href="http://linkedgedata.org/vocabulary#basketball%3Bhockey">http://linkedgedata.org/vocabulary#basketball%3Bhockey</a>
<a href="http://linkedgedata.org/vocabulary#fishing">http://linkedgedata.org/vocabulary#fishing</a>
<a href="http://linkedgedata.org/vocabulary#residential">http://linkedgedata.org/vocabulary#residential</a>
<a href="http://linkedgedata.org/vocabulary#hockey%3Bbasketball%3Blacrosse">http://linkedgedata.org/vocabulary#hockey%3Bbasketball%3Blacrosse</a>
<a href="http://linkedgedata.org/vocabulary#reservoir">http://linkedgedata.org/vocabulary#reservoir</a>
<a href="http://linkedgedata.org/vocabulary#hockey%3Bbasketball">http://linkedgedata.org/vocabulary#hockey%3Bbasketball</a>
<a href="http://linkedgedata.org/vocabulary#traffic_signals">http://linkedgedata.org/vocabulary#traffic_signals</a>
<a href="http://linkedgedata.org/vocabulary#townhall">http://linkedgedata.org/vocabulary#townhall</a>
<a href="http://linkedgedata.org/vocabulary#subway_entrance">http://linkedgedata.org/vocabulary#subway_entrance</a>
<a href="http://linkedgedata.org/vocabulary#bus_station">http://linkedgedata.org/vocabulary#bus_station</a>
<a href="http://linkedgedata.org/vocabulary#DAM">http://linkedgedata.org/vocabulary#DAM</a>
<a href="http://linkedgedata.org/vocabulary#attraction">http://linkedgedata.org/vocabulary#attraction</a>
<a href="http://linkedgedata.org/vocabulary#council+offices">http://linkedgedata.org/vocabulary#council+offices</a>
<a href="http://linkedgedata.org/vocabulary#rail">http://linkedgedata.org/vocabulary#rail</a>
<a href="http://linkedgedata.org/vocabulary#hockey">http://linkedgedata.org/vocabulary#hockey</a>
<a href="http://linkedgedata.org/vocabulary#american+football+%3B+athletics">http://linkedgedata.org/vocabulary#american+football+%3B+athletics</a>
<a href="http://linkedgedata.org/vocabulary#athletics">http://linkedgedata.org/vocabulary#athletics</a>
<a href="http://linkedgedata.org/vocabulary#football%3Bsoccer">http://linkedgedata.org/vocabulary#football%3Bsoccer</a>
<a href="http://linkedgedata.org/vocabulary#sports">http://linkedgedata.org/vocabulary#sports</a>
<a href="http://linkedgedata.org/vocabulary#river">http://linkedgedata.org/vocabulary#river</a>
<a href="http://linkedgedata.org/vocabulary#playground">http://linkedgedata.org/vocabulary#playground</a>
<a href="http://linkedgedata.org/vocabulary#hotel">http://linkedgedata.org/vocabulary#hotel</a>
<a href="http://linkedgedata.org/vocabulary#viewpoint">http://linkedgedata.org/vocabulary#viewpoint</a>
<a href="http://linkedgedata.org/vocabulary#athletics+soccer">http://linkedgedata.org/vocabulary#athletics+soccer</a>
<a href="http://linkedgedata.org/vocabulary#bus_stop">http://linkedgedata.org/vocabulary#bus_stop</a>
<a href="http://linkedgedata.org/vocabulary#sport">http://linkedgedata.org/vocabulary#sport</a>
<a href="http://linkedgedata.org/vocabulary#icehockey">http://linkedgedata.org/vocabulary#icehockey</a>
<a href="http://linkedgedata.org/vocabulary#multi">http://linkedgedata.org/vocabulary#multi</a>
<a href="http://linkedgedata.org/vocabulary#skiing">http://linkedgedata.org/vocabulary#skiing</a>
<a href="http://linkedgedata.org/vocabulary#rugby%3B+athletics">http://linkedgedata.org/vocabulary#rugby%3B+athletics</a>
<a href="http://linkedgedata.org/vocabulary#henge">http://linkedgedata.org/vocabulary#henge</a>
<a href="http://linkedgedata.org/vocabulary#rugby">http://linkedgedata.org/vocabulary#rugby</a>
<a href="http://linkedgedata.org/vocabulary#nature_reserve">http://linkedgedata.org/vocabulary#nature_reserve</a>
<a href="http://linkedgedata.org/vocabulary#airfield">http://linkedgedata.org/vocabulary#airfield</a>
<a href="http://linkedgedata.org/vocabulary#basketball%3Bvolleyball%3Bgymnastics%3B">http://linkedgedata.org/vocabulary#basketball%3Bvolleyball%3Bgymnastics%3B</a>
<a href="http://linkedgedata.org/vocabulary#construction">http://linkedgedata.org/vocabulary#construction</a>
<a href="http://linkedgedata.org/vocabulary#drain">http://linkedgedata.org/vocabulary#drain</a>

<a href="http://linkedgedata.org/vocabulary#plan.at%3Ab#%3Ariver">http://linkedgedata.org/vocabulary#plan.at%3Ab#%3Ariver</a>
<a href="http://linkedgedata.org/vocabulary#canal">http://linkedgedata.org/vocabulary#canal</a>
<a href="http://linkedgedata.org/vocabulary#trekking">http://linkedgedata.org/vocabulary#trekking</a>
<a href="http://linkedgedata.org/vocabulary#_water_fixme_afterwards">http://linkedgedata.org/vocabulary#_water_fixme_afterwards</a>
<a href="http://linkedgedata.org/vocabulary#athletics%3Bsoccer%3Blacrosse">http://linkedgedata.org/vocabulary#athletics%3Bsoccer%3Blacrosse</a>
<a href="http://linkedgedata.org/vocabulary#industrial">http://linkedgedata.org/vocabulary#industrial</a>
<a href="http://linkedgedata.org/vocabulary#true">http://linkedgedata.org/vocabulary#true</a>
<a href="http://linkedgedata.org/vocabulary#park">http://linkedgedata.org/vocabulary#park</a>
<a href="http://linkedgedata.org/vocabulary#land">http://linkedgedata.org/vocabulary#land</a>
<a href="http://linkedgedata.org/vocabulary#railway">http://linkedgedata.org/vocabulary#railway</a>
<a href="http://linkedgedata.org/vocabulary#climbing">http://linkedgedata.org/vocabulary#climbing</a>
<a href="http://linkedgedata.org/vocabulary#town_hall">http://linkedgedata.org/vocabulary#town_hall</a>
<a href="http://linkedgedata.org/vocabulary#train_station">http://linkedgedata.org/vocabulary#train_station</a>
<a href="http://linkedgedata.org/vocabulary#undefined">http://linkedgedata.org/vocabulary#undefined</a>
<a href="http://linkedgedata.org/vocabulary#tower">http://linkedgedata.org/vocabulary#tower</a>
<a href="http://linkedgedata.org/vocabulary#Coffs+Harbour">http://linkedgedata.org/vocabulary#Coffs+Harbour</a>
<a href="http://linkedgedata.org/vocabulary#recreation_ground">http://linkedgedata.org/vocabulary#recreation_ground</a>
<a href="http://linkedgedata.org/vocabulary#athletics%3B+soccer">http://linkedgedata.org/vocabulary#athletics%3B+soccer</a>
<a href="http://linkedgedata.org/vocabulary#basketball%3B+hockey">http://linkedgedata.org/vocabulary#basketball%3B+hockey</a>
<a href="http://linkedgedata.org/vocabulary#Rapids">http://linkedgedata.org/vocabulary#Rapids</a>
<a href="http://linkedgedata.org/vocabulary#archaeological_site">http://linkedgedata.org/vocabulary#archaeological_site</a>
<a href="http://linkedgedata.org/vocabulary#mini_roundabout">http://linkedgedata.org/vocabulary#mini_roundabout</a>
<a href="http://linkedgedata.org/vocabulary#Baseball">http://linkedgedata.org/vocabulary#Baseball</a>
<a href="http://linkedgedata.org/vocabulary#post_box">http://linkedgedata.org/vocabulary#post_box</a>
<a href="http://linkedgedata.org/vocabulary#tertiary">http://linkedgedata.org/vocabulary#tertiary</a>
<a href="http://linkedgedata.org/vocabulary#survey_point">http://linkedgedata.org/vocabulary#survey_point</a>
<a href="http://linkedgedata.org/vocabulary#restaurant">http://linkedgedata.org/vocabulary#restaurant</a>
<a href="http://linkedgedata.org/vocabulary#political">http://linkedgedata.org/vocabulary#political</a>
<a href="http://linkedgedata.org/vocabulary#reservoir_covered">http://linkedgedata.org/vocabulary#reservoir_covered</a>
<a href="http://linkedgedata.org/vocabulary#pedestrian">http://linkedgedata.org/vocabulary#pedestrian</a>
<a href="http://linkedgedata.org/vocabulary#golf_course">http://linkedgedata.org/vocabulary#golf_course</a>
<a href="http://linkedgedata.org/vocabulary#American+football">http://linkedgedata.org/vocabulary#American+football</a>
<a href="http://linkedgedata.org/vocabulary#famous_school">http://linkedgedata.org/vocabulary#famous_school</a>
<a href="http://linkedgedata.org/vocabulary#cycling">http://linkedgedata.org/vocabulary#cycling</a>
<a href="http://linkedgedata.org/vocabulary#aussie_rules">http://linkedgedata.org/vocabulary#aussie_rules</a>
<a href="http://linkedgedata.org/vocabulary#cricket">http://linkedgedata.org/vocabulary#cricket</a>
<a href="http://linkedgedata.org/vocabulary#commercial">http://linkedgedata.org/vocabulary#commercial</a>
<a href="http://linkedgedata.org/vocabulary#theatre">http://linkedgedata.org/vocabulary#theatre</a>

<a href="http://linkedgeodata.org/vocabulary#museum">http://linkedgeodata.org/vocabulary#museum</a>
<a href="http://linkedgeodata.org/vocabulary#city_wall">http://linkedgeodata.org/vocabulary#city_wall</a>
<a href="http://linkedgeodata.org/vocabulary#public_building">http://linkedgeodata.org/vocabulary#public_building</a>
<a href="http://linkedgeodata.org/vocabulary#disused">http://linkedgeodata.org/vocabulary#disused</a>
<a href="http://linkedgeodata.org/vocabulary#castle">http://linkedgeodata.org/vocabulary#castle</a>
<a href="http://linkedgeodata.org/vocabulary#locality">http://linkedgeodata.org/vocabulary#locality</a>
<a href="http://linkedgeodata.org/vocabulary#australian_football">http://linkedgeodata.org/vocabulary#australian_football</a>
<a href="http://linkedgeodata.org/vocabulary#derelict_canal">http://linkedgeodata.org/vocabulary#derelict_canal</a>
<a href="http://linkedgeodata.org/vocabulary#picnic_site">http://linkedgeodata.org/vocabulary#picnic_site</a>
<a href="http://linkedgeodata.org/vocabulary#Arena">http://linkedgeodata.org/vocabulary#Arena</a>
<a href="http://linkedgeodata.org/vocabulary#narrow_gauge">http://linkedgeodata.org/vocabulary#narrow_gauge</a>
<a href="http://linkedgeodata.org/vocabulary#plan.at%3Aami%3Ariver">http://linkedgeodata.org/vocabulary#plan.at%3Aami%3Ariver</a>
<a href="http://linkedgeodata.org/vocabulary#monument">http://linkedgeodata.org/vocabulary#monument</a>
<a href="http://linkedgeodata.org/vocabulary#caravan_site">http://linkedgeodata.org/vocabulary#caravan_site</a>
<a href="http://linkedgeodata.org/vocabulary#manmade_lake">http://linkedgeodata.org/vocabulary#manmade_lake</a>
<a href="http://linkedgeodata.org/vocabulary#rugby_league">http://linkedgeodata.org/vocabulary#rugby_league</a>
<a href="http://linkedgeodata.org/vocabulary#rugby%3Bfootball">http://linkedgeodata.org/vocabulary#rugby%3Bfootball</a>
<a href="http://linkedgeodata.org/vocabulary#football%3Bsoccer%3Brugby+league">http://linkedgeodata.org/vocabulary#football%3Bsoccer%3Brugby+league</a>
<a href="http://linkedgeodata.org/vocabulary#sailing">http://linkedgeodata.org/vocabulary#sailing</a>
<a href="http://linkedgeodata.org/vocabulary#multi%3Brugby%3Bathletics">http://linkedgeodata.org/vocabulary#multi%3Brugby%3Bathletics</a>
<a href="http://linkedgeodata.org/vocabulary#military">http://linkedgeodata.org/vocabulary#military</a>
<a href="http://linkedgeodata.org/vocabulary#football+cricket">http://linkedgeodata.org/vocabulary#football+cricket</a>