

Semantic enrichment of Volunteered Geographic Information using Linked Data: a use case scenario for disaster management

Master of Science Thesis

Stanislav Ronzhin

July 1st 2015

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



Acknowledgements

Foremost, I would like to express my sincere gratitude to my supervisor dr.ir. R.L.G. (Rob) Lemmens for the help and guidance during the work. Thanks to my wife Tatiana, who stood up to all of my countless talks about the thesis. Thanks to my daughter Leela, watching her development motivated my work and gave precious inspiration. I am also grateful to Miyusova Natalia Pavlovna, my geography schoolteacher, who inspired me with the love of answering where questions.

Abstract

Web 2.0 data deluge provoked by the development of collaborative tools has affected numerous domains. In the context of the crowdsourcing of geographic information, the concept of Volunteered Geographic Information (VGI) has emerged. However, the quality and usability of VGI is a subject of a debate. Data often comes unstructured with unknown accuracy and lacking reliability. Semantic integration of VGI with relevant entities in the Linked Open Data (LOD) cloud has been seen as a remedy to overcome weakness of a crowdsourced data. The LOD cloud makes it possible to semantically enrich unstructured user-generated content with structured information presented in the LOD resources.

This thesis questions ***to what extent the Linked Open Data cloud can help to semantically enrich volunteered geographic information in order to better answer queries in the context of crisis and disaster relief operations.*** Data produced by the Ushahidi project during the Chilean earthquake of 2011 has been chosen as an example of a disaster related VGI.

In general, the work implied a construction of the proof of concept. The first two steps have included a conversion of the data into the Resource Description Framework (RDF) using vocabularies and establishing of semantic links to relevant LOD entities. The use of the Management of a Crisis vocabulary has increased semantic interoperability of the original data. Semantic enrichment achieved via established links has helped to overcome ambiguous georeferencing of the data thus allowing a robust spatial dimension to the data. Emerged spatial capabilities made it possible to access data entities using spatial queries. In turn, the latter provided a straightforward mechanism for data retrieval, for instance, from DBpedia.

The work has shown that the LOD cloud can be perceived as a giant informational skeleton. Scattered and disconnected blobs of unstructured data, being attached to this skeleton, acquire an integrated dataspace where standardized methods of data access and manipulation can be used. Despite of the fact, the work dealt with the disaster-related VGI, the demonstrated approach can be applied to any VGI.

Table of Contents

Acknowledgements.....	ii
Abstract.....	iii
Index of figures.....	vi
Index of listings.....	vii
List of abbreviations.....	viii
Chapter 1. Introduction.....	1
1.1 Background.....	1
1.2 Use case scenario rationale and motivation.....	3
1.3 Problem statement.....	4
1.4 Research objectives and questions.....	4
1.4.1 Sub-objective one -To integrate disaster VGI into the LOD cloud.....	5
1.4.2 Sub-objective two - To evaluate methods for the construction of semantic queries ...	5
1.4.3 Sub-objective three - To evaluate the results.....	6
Chapter 2. Linked Data and its applications.....	7
2.1 Semantic Web and Linked Data technologies.....	7
2.2 Linked Open Data cloud and its geospatial content.....	10
2.3 Integration of data into the LOD cloud.....	11
2.4 Tools for Integration of data into the LOD cloud.....	13
2.4.1 Conversion tools and RDF generators.....	13
2.4.2 RDF validators.....	13
2.4.3 Semantic Web browsers.....	14
2.4.4 Triplestores.....	15
2.4.5 Discovery and establishing of semantic links between data sets.....	15
2.4.6 Query builders and constructors for SPARQL.....	16
2.5 Visualization of RDF data and SPARQL result set.....	16
Chapter 3. Information in Emergency Management.....	18
3.1 Information management in Emergency Management.....	18
3.2 User Generated Content and crowdsourcing in Emergency Management.....	19
3.3 Linked Data for Emergency Management.....	20
Chapter 4. Prototyping the case study.....	23
4.1 Data.....	23
4.2 Work package 1. Conversion of data into RDF with links to LinkedGeoData.....	25
4.3 Work Package 2. Construction of queries for a semantic enrichment.....	32
Chapter 5. Evaluation of the results.....	40
5.1 Evaluation of work package 1.....	40
5.2 Evaluation of work package 2.....	41
5.3 Evaluation of emerged data management techniques.....	43
Chapter 6. Discussion, Conclusions and Recommendations.....	48
6.1 Discussion.....	48
6.2 Conclusions.....	49
6.2.1 Main conclusion.....	49
6.2.2 Answering sub research questions.....	50
6.2 Recommendations.....	53
References.....	56
Appendices.....	62
Appendix A. Example of Ushahidi data.....	62
Appendix B. Ontology mapping between the Ushahidi categories and MOAC for Chile.....	63

Appendix C. Ontology mapping between the Ushahidi categories and MOAC for Haiti..... 65
Appendix D. List of missing LGD objects 68
Appendix E. Table of namespace prefixes 69
Appendix F. List of the selected reports..... 70

Index of figures

Figure 1. Linking Open Data cloud diagram as of August 2014.	2
Figure 2. Linked Open Vocabularies classification	12
Figure 3. Information Management at the core of Humanitarian Decision Making Process for all the Clusters throughout Crisis/Disaster Management phases (Credit: UNOCHA)	19
Figure 4. Workflow diagram for the first work package.	25
Figure 5. Graph visualization of the report 4349 in the initial state (A) after the conversion (B)	28
Figure 6. Example five reports assigned with a category "Collapsed Structure"	29
Figure 7. Results retrieved by the query in Listing 4.	30
Figure 8. Georeferencing of the report 4349.....	31
Figure 9. Workflow diagram for the second work package.	33
Figure 10. Representation of Hospital de Coronel in LinkedGeoData.	33
Figure 11. Computational environment of the proof of concept.	34
Figure 12. Example results retrieved by the query in Listing 7.	35
Figure 13. Description of Lota on DBpedia.	36
Figure 14. 2 km proximity to Hospital de Lota.	37
Figure 15. Results retrieved by the query in Listing 13.....	41
Figure 16. Interface of Chile Ushahidi deployment.	43
Figure 17. Example results retrieved by the query in Listing 15.....	44

Index of listings

Listing 1. SPARQL query to retrieve plants from the linkedplants database	9
Listing 2. RDF-based representation of report 4349.....	27
Listing 3. The first test query.....	29
Listing 4. Selection of the reports with more than 2 categories.....	29
Listing 5. Report 3790 with 10 assigned categories.....	30
Listing 6. Final version of the report 4349.	32
Listing 7. SPARQL query for selection of the triples related to a LGD object.	34
Listing 8. INSERT query.....	35
Listing 9. INSERT query selecting geometries	35
Listing 10. Selection of DBpedia entries about cities in 2 km proximity to a report	37
Listing 11. Query with OPTIONAL keywords	38
Listing 12. INSERT query for DBpedia triples	39
Listing 13. Query to check a datatype of geometry representation.....	41
Listing 14. Query to change the datatype of a geometry representation.	42
Listing 15. Selection of blocked roads and their geometries.....	44
Listing 16. Selection of bridges located in 10-km proximity to compromised bridges.....	45
Listing 17. Selection of officials of populated places where operating hospitals are located.	45
Listing 18. Prioritizing of the reports based on the density of population.	46

List of abbreviations

3W	Who What Where
AJAX	asynchronous JavaScript and XML
API	Application Program Interface
ATM	Automated Teller Machine
BSD	Berkeley Software Distribution
CSV	Comma-Separated Values
DAML	DARPA Agent Markup Language
DBMS	Database Management System
EM	Emergency Management
EU	European Union
FP7	7 th Framework Programme for Research and Technological Development
GiST	Generalized Search Tree
GML	Geography Markup Language
GUI	Graphical User Interface
HIC	Humanitarian Information Centers
HLX	Humanitarian eXchange Language
HTML	Hyper Text Markup Language
HTTP	Hyper Text Transfer Protocol
IASC	Inter-Agency Standing Committee
IM	Information Management
IMU	Information Management Unit
IRI	International Resource Identifier
IRIN	Integrated Regional Information Networks
ISO	International Organization for Standardization
IEC	International Electrotechnical Commission
IT	Information Technology
JSON	JavaScript Object Notation
kNN	k-Nnearest Neighbor
LGD	LinkedGeoData
LOD	Linked Open Data
LOV	Linked Open Vocabulary
MOAC	Management of a Crisis
NGO	Non-Governmental Organization
OGC	Open Geospatial Consortium
OSM	OpenStreetMap
OWL	Web Ontology Language
PHP	PHP: Hypertext Preprocessor
RCC8	Region connection calculus
RDB	Relational Database
RDF	Resource Description Framework
RDFa	Resource Description Framework in Attributes
RDFS	Resource Description Framework Schema
SIOC	Semantically-Interlinked Online Communities
SMS	Short Message Service
SPARQL	SPARQL Protocol and RDF Query Language

UN	United Nations
UMBEL	Upper Mapping and Binding Exchange Layer
UNOCHA	The United Nations' Office for the Coordination of Humanitarian Affairs
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
USGS	United States Geological Survey
UTC	Coordinated Universal Time
VGI	Volunteered Geographic Information
W3C	World Wide Web Consortium
WGS	World Geodetic System
WKT	Well-Known Text
XML	eXtensible Markup Language
YAGO	Yet Another Great Ontology

Chapter 1. Introduction

1.1 Background

Deluge of crowdsourced data

The emergence of Web 2.0 has led to a data explosion, initially in non-spatial information and subsequently also in the geographic domain (Ballatore et al., 2013). This explosion occurred in the wake of inventions that enabled users to increase their participation in the content creation. The term “neogeography” has been coined to encompass this rapid and complex generation of technological and social practices aimed at collection and exploitation of geo-referenced information, using collaborative web tools (Turner 2006, O'reilly, 2007). Goodchild named the crowdsourcing of geographic information as Volunteered Geographic Information (VGI). This term emphasizes the participatory nature of data production using voluntary actions. Literally, everyone with an Internet connection can act as a “sensor”, therefore facilitating the collection and maintaining of geographical information. (Goodchild, 2007).

VGI projects have already acquired a giant amount of geo-referenced information. For instance, as of May 2014, Wikimapia contained more than 23 million objects marked by registered users and guests (Wikimapia, 2014). More than 4.5 million articles in Wikipedia had geotags in August 2014 (Geographic intersections of languages in Wikipedia, 2014). However, these numbers are almost nothing in comparison with OpenStreetMap project, where around 1.7 million users created more than 2.5 billion nodes as of October 2014 (OpenStreetMap, 2014). Thus, crowdsourcing initiatives have become an integral part of the Geospatial Web. The latter is yet another neologism of Web 2.0 that refers to “the use of the internet to deliver geographic information and maps” (Haklay et al., 2008).

However, despite of the fact that crowdsourcing initiatives generate vast amount of information, the quality and usability of the content is a subject of debate. On the one hand, data coverage is not complete or consistent across the globe; areas with higher population density receive more attention from users than less populated territories (Graham et al., 2014). In addition, in many cases people collect information without any guidance or instructions. As a result, the accuracy of such data is often unknown, as there are no systemic and comprehensive quality assurance processes integral to the data collection (Haklay et al., 2008). On the other hand, most of the times, data collection is facilitated by dilettantes with different professional and educational background. This leads to inconsistent information (Goodchild, 2008); different people can categorize the same phenomenon differently. Crowdsourced data often comes unstructured, in different formats and of heterogeneous reliability, which makes the integration of such data sets to be far from trivial.

Semantic Web to the rescue

The spectacular growth of unstructured information collected by users online prompted Tim Berners-Lee to envisage the advent of the so-called Semantic Web (Berners-Lee et al, 2001). He defines the Semantic Web as "a web of data that can be processed directly and indirectly by machines"(Berners-Lee et al, 2001). In this concept, the Internet 'understands' the pieces of information it stores and is able to make logical connections between them. This allows people who put individual items of data on the Web to link them with other pieces of data. Data semantics are expressed in subject-predicate-object triples encoded in languages such as RDF (Resource Description Framework). These triples constitute large collections of statements about real world entities. The concept of the Semantic Web was further promoted through the Linked Data initiative. The latter refers to the recommended best practices for exposing, sharing, and connecting RDF data via

dereferenceable URIs on the Semantic Web. Links between URIs glue datasets together allowing them to become parts of a single global data set.

The Linking Open Data project, a grassroots community effort founded in January 2007 and supported by the W3C Semantic Web Education and Outreach Group, aims to bootstrap the Web of Data by identifying existing data sets that are available under open licenses (Bizer et al., 2009). Figure 1 illustrates the constellation of resources available in the Linked Open Data (LOD) cloud. As can be seen from the Figure, numerous crowdsourcing initiatives have already undertaken a task of publishing their data in compliance with the Linked Data principles. Examples include DBpedia (Wikipedia content published in RDF) GeoNames, LinkedGeoData (RDF version of OpenStreetMap data) and many others. All together, they constitute a unique source of knowledge from numerous domains.

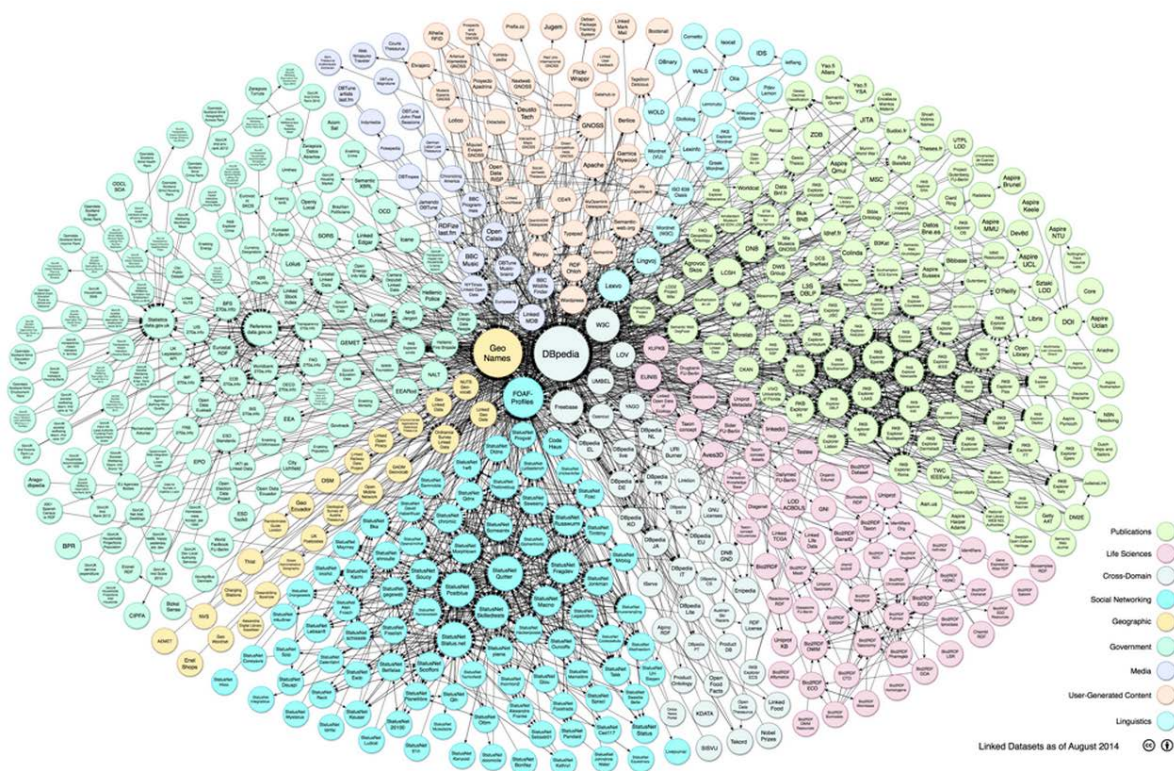


Figure 1. Linking Open Data cloud diagram as of August 2014.

More data, more understanding

Janowicz et al (2012) promotes the idea that semantic integration will allow researchers to combine data from heterogeneous sources to gain a more holistic understanding of places. In other words, the interlinking of user-generated content with relevant nodes in the LOD cloud makes it possible to reveal additional relationships between data entities that would be hidden or implicit if we use the original data set alone. In this way, linked data can be used as a source of comprehensive background knowledge for interpretation of geographical information. Integration of a dataset into the LOD cloud semantically enriches the original content of the data, providing additional meaning and allowing users to answer questions that are more complex. Feliachi (2013) shows that semantic integration of two data sets would enable users to take advantage of both information sources. In the case when a VGI data set is integrated with the LOD cloud we can literally take advantage not only of resources which are directly interlinked with that VGI dataset, but literally of all the knowledge

presented in the LOD cloud, due to the great interconnectivity between data sets published as linked data.

However, as Goodwin et al (2008) notice, although technically establishing links between data may be trivial, semantically it can be hard. The challenge is to make a trade-off between two main design considerations; links must be semantically accurate and must lead to as many external RDF nodes as possible. Another challenge comes from the complexity of information retrieval from the LOD cloud. Structure of the Linked Open Data cloud is complex and heterogeneous. Some of the resources are domain specific (e.g Greek administrative geography) when others contain cross-domain ontologies and information (e.g DBpedia). In order to construct a proper query, a user has to be aware of particular vocabularies used in an informational resource. In addition, geographic information (e.g. VGI) features a unique informational component, namely space. This spatial component requires distinctive computational approaches and specific data models and ontologies to store spatial features. Only relatively recent developments (Battle & Kolas, 2012) have provided spatial extensions to the semantic technologies enabling computationally efficient querying and retrieval of geographic information stored in RDF.

This thesis will investigate and discuss the extent of possible semantic enrichment of VGI achieved by integrating it into the LOD cloud. Obstacles and solutions will be identified from the construction of proof of concept for a use case scenario from a crisis and disaster management practice. The motivation and rationale of the use case scenario will be explained in the next section.

1.2 Use case scenario rationale and motivation

Undoubtedly, within the dynamic context of humanitarian operations, the availability of timely, relevant and reliable information is one of the crucial factors influencing the success of relief actions. This need is widely recognized by the humanitarian organizations (Van de Walle et al., 2009). For instance, Sir John Holmes, The UN Emergency Relief Coordinator, puts it very straightforward “information is very directly about saving lives. If we take the wrong decisions (...), because our knowledge is poor, we are condemning some of the most deserving to death or destitution” (Haggarty & Naidoo, 2008). In a major emergency, humanitarian agencies put massive combined effort to collect, synthesize and analyze situational data. This is especially the case during the first hours after a catastrophe, when rapid assessments to estimate the needs of affected population take place.

Web 2.0 technologies have provided a number of tools that help to gather, to process and to map pleas for help. Members of communities in a disaster struck area send help requests using their mobile phones or computers connected to the Web. Ushahidi and Sahana Eden are the two most common examples of open-source platforms that allow reporting on crisis related events via numerous media channels, including SMS, Email, mobile application and via the website. It helps disaster managers, emergency response practitioners to track users’ reports on a map and timeline, know the needs of the affected victims and coordinate emergency agencies and aid resources (Duc et al., 2014). Ushahidi helped to process about 40 000 reports received from affected population during Haiti earthquake in 2010. Sahana Eden was developed by the information technology (IT) community in Sri Lanka in order to assist the country recover after the earthquake and tsunami in 2004 (Duc et al., 2014).

However, despite of the fact that both platforms provide near-real time information about needs of people in a disaster-struck area, utilization of this information has not reached its full potential. On the one hand, lacking semantic interoperability between user-generated reports and official data sources used by disaster relief organizations significantly hampers the integration of data (Ortman et al., 2011). On the other hand, agencies have limited staff and resources to harness this massive stream of user generated content even though it contains pieces of highly relevant - but unknown to the

decision makers - information (Goodchild & Glennon, 2010, Schulz et al., 2012). The size of crowdsourced data is a way too big to be efficiently handled by the humanitarian agencies.

Therefore, organizations are limited to take advantage of resources such as OpenStreetMap.org. However, harvesting disaster related information from multiple data sources across the web would contribute to a better situational awareness and operational picture. What if there is enough information in the LOD cloud to satisfy the needs of decision makers? For example, consider a combined data source where help requests are integrated with OpenStreetMap data and information about hospitals in the region. Such a combination would give an answer to the question “How to reach the closest operating hospital avoiding road blockages and who is in charge at that place?” Another example of a highly relevant question for emergency management staff is “Can we cross the bridge with a 12 ton fire truck?” OpenStreetMap has a number of tags with key:value pairs describing a legal weight and height limit for using a road or a bridge. Therefore, this question can be answered using a combination of data collected by Ushahidi or Sahana Eden and OpenStreetMap.

1.3 Problem statement

Put differently, Linked Data is a collection of typed links between data from heterogeneous sources. These links are machine-readable and their meaning is explicitly defined. This allows users to answer complex queries spanning multiple, heterogeneous data sources from different scientific domains.

In respect to crisis and disaster response, integration of disaster related crowdsourced information with information available on the web of data using the Linked Data principles seems to be very promising. Several works (Borges et al., 2011, Heim et al., 2011, Mijovic et al., 2013, Ortman et al., 2011, Schulz et al., 2012) approach harnessing of Linked Data for the purpose of emergency management and response. However, none of them elaborate to what extent semantic integration of disaster related VGI with relevant entities in the LOD cloud can help to answer relevant questions for relief operations. By transforming help requests collected by crowdsourcing platforms into linked RDF triples, the initial semantics of these VGI is enriched with the content from almost the entire LOD cloud. However, how significant this semantic enrichment is and what kind of relationships based on this enriched data can be revealed still remain questions. In addition, invention of GeoSPARQL and spatial indexing in triple stores have enabled spatial reasoning for SPARQL queries posed across linked data resources. In turn, this considerably increases the range of possible relationships between data entities to be discovered.

1.4 Research objectives and questions

Semantic integration of disaster related volunteered geographic information with relevant entities in the Linked Open Data (LOD) cloud, semantically enriches the content of crowdsourced testimonies. Background knowledge from the LOD cloud provides underlying meaning for help requests collected via the Ushahidi platform. This approach allows answering questions, which are highly relevant to crisis and disaster relief operations but cannot be easily answered using traditional information management techniques adopted by domain experts. Therefore, ***the main research question of this study is to analyze to what extent the Linked Open Data cloud can help to semantically enrich volunteered geographic information in order to better answer queries in the context of crisis and disaster relief operations.***

The main objective is decomposed into three sub-objectives with relevant research questions as follows:

1.4.1 Sub-objective one -To integrate disaster VGI into the LOD cloud

1. *What standards and tools facilitate semantic integration of disaster related crowdsourced information?* By this question, the technical component of the semantic integration is examined. It touches upon the working mechanism of the Semantic Web and Linked Data as well as overviews the state-of-the art in this field.
2. *What ontologies can be used for data conversion into RDF and how?* So far, numerous ontologies supporting integration of different data sources have been developed. In the case of ontologies for disaster related VGI, there are not so many options. (Liu et al., 2013). Ortmann et al (2011) present one of the example ontologies. They develop Management of a Crisis (MOAC) Vocabulary that provides means for interlinking data across traditional humanitarian agencies, crowdsourced volunteered technical committees and affected populations. However, applicability of this vocabulary for integration of crowdsourced data into the Linked Open Data cloud can be questioned. In addition, what if less domain specific ontologies are able to support integration of disaster related VGI into the Linked Open Data cloud?
3. *What Linked Data Hubs can be used for establishing outgoing links from RDF-based crowdsourced data?* The structure of the Linked Open Data cloud is complex and heterogeneous. Some of the resources are domain specific (e.g Greek administrative geography) where others contain cross-domain ontologies and information (e.g DBpedia). In addition, geospatial resources can be used for link generation based on location; in contrast, non-spatial resources provide textual information that can serve as a source of comprehensive background knowledge.
4. *What is the difference between integration of VGI into the Linked Open Data cloud in the case of poor and rich information environment?* Richness of information coverage provided by Linked Data resources available on the Web vary across the globe. Graham et al (2014) reveal uneven geographical distribution of Wikipedia content. In turn, such a distribution influences the number of RDF nodes for integration. This question provides details on obstacles associated with interlinking of user generated messages and Linked Data resources in different parts of the Earth.

1.4.2 Sub-objective two - To evaluate methods for the construction of semantic queries

5. *What questions are difficult to answer using existing information management techniques in the context of crisis and disaster management?* By this question, the insight into up-to-date information management techniques used for crisis and disaster management is gained. Several works (e.g. Schulz et al., 2012, Ortmann et al., 2011) examine drawbacks in information management techniques adopted by emergency relief organizations. Extended literature review will further investigate what limitations in data management affect integration of heterogeneous data sources during a crisis and disaster relief.
6. *Which of them can be answered by posing GeoSPARQL queries against an RDF-based disaster related VGI linked to multiple Linked Data resources?* Ortmann et al., 2011 proposed the use of Linked Data technologies for better handling of crowdsourced information during a catastrophic event. This sub-question explains how identified limitations of existing information management techniques can be overcome using spatially enabled semantic technologies. Based on this, particular solutions using GeoSPARQL will be developed for known difficulties.
7. *How to construct a SPARQL query? What tools are able to assist in the construction of a SPARQL query for a SPARQL endpoint?* SPARQL is an RDF query language, able to retrieve and manipulate data stored in Resource Description Framework. This sub question goes into

details on SPARQL technologies and investigates how to create a query in this language. Since 2008, when the standard was issued, several tools have been developed to facilitate an automatic or semi-automatically construction of queries. One of the example is ViziQuer (Zviedris & Barzdins, 2011); a software that helps to create SPARQL queries and explore SPARQL endpoints. Tabulator (Berners-Lee et al., 2006) is another example. However, the range of available tools is not limit by these two examples. Other options will be explored and evaluated.

8. *What is special about spatial SPARQL queries?* This question investigates peculiarities associated with the use of GeoSPARQL, a spatially-enabled version of SPARQL query language. Implementation of spatial indexing in triple stores together with development of a well-understood ontology for representation of spatial objects allowed spatial reasoning in SPARQL queries. This made it possible to efficiently answer a query such as "What are all the schools near Atlanta, GA that are within 100 meters of a railway?" addressed to just two Linked Data resources – GeoNames and USGS rail lines dataset (Battle & Kolas, 2012). The current sub question also touches upon issues such as "How to add a spatial condition into a SPARQL query?" and 'What tools can provide required for this functionality?

1.4.3 Sub-objective three - To evaluate the results

9. *What tools can provide visualization of the results retrieved by GeoSPARQL queries?* Visualization plays an important role in the verification of results. It gives ideas about mutual location of retrieved objects.
10. *How correct are the results retrieved by GeoSPARQL queries?* The correctness of the results will be evaluated based on information available on the Web that are different from information obtained from the Linked Open Data cloud.
11. *To what extent does difference in richness of informational environment influence the robustness of the retrieved results?* This question further elaborates the influence of uneven geographical distribution of information presented in the Linked Open Data cloud on the results retrieved by GeoSPARQL queries.

Chapter 2. Linked Data and its applications

By this chapter the overview of the Linked Data technologies and their implementations are given. The first subsection introduces the notion of the Semantic Web and explains the principles of Linked Data. Due to the fact that this thesis deals with geospatial data, the geo- component of the Linked Data Cloud is thoroughly investigated in the second subsection. Then, subsections three and four elaborate on how to integrate data into the Linked Data Cloud and which tools can help in this process. The last subsection focuses on the visualization of Linked Data presenting tools that make exploration of Linked Data resources more human friendly.

2.1 Semantic Web and Linked Data technologies

Historically, the explosive growth of the Web has been driven by two factors (Heath & Bizer, 2011). First, people are free to publish any documents they want on the Web without registration of them in any register. Once published, a document can be immediately accessed by any user with a web browser. Second, to make these documents searchable, search engines crawl the web discovering new documents and provide links to them by a request. However, although this approach provides a very straightforward mechanism for making documents accessible online, it significantly limits possible retrieval methods to keyword searches or matches of sub-strings. The main shortcoming of such an approach lies in the inability of search engines to support more complex data structures than flat text strings. Therefore, higher-level computational operations that require querying, analysis, comparison, combination or integration of data are not possible due to the lack of methods that make compatible information available (Egenhofer, 2002).

Principles of Linked Data

The idea to bring more human-like reasoning into the process of data retrieval as well as to make machines understand pieces of information they store motivated Tim Berners-Lee to begin an effort to investigate foundations for the next stage of the Web, called the Semantic Web (Berners-Lee et al., 2001). The concept of semantic web formulated by Tim Berners-Lee is "a web of data that can be processed directly and indirectly by machines". The first step towards this new Web of Data was the introduction of a set of best practices for publishing and interlinking structured data on the Web that became known as the Linked Data principles (Berners-Lee, 2006). These principles are as follows:

1. Use URIs as names for things.
2. Use HTTP URIs, so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL).
4. Include links to other URIs, so that they can discover more things.

In order to better understand the core idea of the Semantic Web and the Linked Data principles the standards and technologies should be explained first.

The first two principles touch upon the use of HTTP URIs to name things. In the concept of the Semantic Web HTTP URIs are used as names for real-world objects and abstract concepts rather than as addresses for Web documents. The content of a data set is described using a simple graph-based data model– the Resource Description Framework (RDF) (Klyne and Carroll, 2004). There are several serializations of RDF, but XML-based (Becket, 2004) and RDFa (Adida & Birbeck, 2008) are two most common formats (Heath & Bizer, 2011). In RDF, a resource is described as a set of statements called triples. A triple represents the basic structure of a simple sentence consisting of three parts, namely a subject, predicate and an object. These three parts can be expressed as URIs, but objects can also be literal. In general, the subject defines the described resource, the predicate, in the middle, shows what

kind of relation exists between subject and object and object is another resource that has a relation with subject.

Vocabularies are meant to provide collections of URIs that can be used as predicates to represent information about a certain domain. One can notice that due to the great variety of domains, a range of possible relations can approach infinity. However, if a suitable term already exists in one of the vocabularies, it should be reused, rather than reinvented. Such an approach maximizes the probability that data can be consumed by applications that may be tuned to well-known vocabularies (Heath & Bizer, 2011).

Ontology vs. Vocabulary vs. Folksonomy

Vocabulary and ontology are two terms that are often used interchangeably. Both notions refer to the formal representation of concepts and relationships used to describe an area of interest. W3C community points out that there is no clear division between what is referred to as “vocabularies” and “ontologies” in the context of the Semantic Web. However, in practice, the word “ontology” is used for more complex, and possibly quite formal collection of terms. In the case when such strict formalism is not necessarily used the trend is to use the term “vocabulary” (W3C, 2014a). In order to avoid possible confusion, both words are used as synonyms in this thesis.

Ontologies can be divided into two classes according to their nature. The first class consists of formal and controlled ontologies. These ontologies are created and maintained by experts as a result of conceptualization and formalization of a domain knowledge (Guarino, 1998). Formal ontologies possess explicit hierarchical structures and can be seen as a layered pyramid with more general concepts situated on the top. Ontologies of the second class, in contrast, are elicited from the knowledge gained as a result of aggregation of user generated tags (Gruber, 2007). Emergence of Web 2.0 allowed users to “tag” with keywords the content they created or encountered. This “collective intelligence” is called “folksonomy”. Vocabularies derived from folksonomies are shallow, lacking hierarchy, or parent-child relation between entities. Tags are literally “equal” to each other, thus, it is impossible to establish any hierarchical structure between categories. Here comes the main difference between top-down approach of formal ontologies and bottom-up approach of folksonomies.

Knowledge from a particular domain or area of concern captured in a vocabulary (or in an ontology) is expressed with the help of a family of knowledge representation languages, namely The Web Ontology Language (OWL). In general, OWL is a semantic markup language for publishing and sharing ontologies on the World Wide Web (Dean et al., 2004). It was originally developed as a vocabulary extension of RDF and was serialized as RDF/XML. W3C OWL Working Group published the first version of OWL in 2003, the standard had matured to the current 2.0 version by 2009.

SPARQL Protocol and Query Language

In the Semantic Web, querying and retrieval of data are facilitated by SPARQL, which is a recursive acronym that stands for **SPARQL Protocol and RDF Query Language**. It allows posing queries against RDF knowledge bases exposed through SPARQL endpoints on the web. In a nutshell, SPARQL queries are based on triple patterns, similar to RDF triples, except that one or more of the constituent resource references are variables (W3C, 2014b). As a query language, it possesses a full set of analytic query operations such as JOIN, SORT, and AGGREGATE. A SPARQL query comprises of a prefix declaration, a dataset definition, a result clause, a query pattern, and query modifiers. Listing 1 gives an example of a simple query that retrieves description of all plants available in the database located at <http://www.linkedplants.com/data/plants>. SPARQL is not only a query language, but also a protocol, and it returns results in a variety of formats: XML, JSON, RDF, and HTML.

SPARQL language specifies four different query forms for different purposes. They are as follows:

- SELECT query extracts raw values from a SPARQL endpoint. The retrieved data are returned in a table format.
- CONSTRUCT query is used to extract information from the SPARQL endpoint. This type of queries returns not only the value of the queried variable but a valid RDF graph with values instead of variables.
- ASK query is used to provide a simple True/False result for a query on a SPARQL endpoint. This query form resembles human YES/NO questions.
- DESCRIBE query form is used to extract an RDF graph containing RDF data that describes the queried resources. It is up to the SPARQL service to choose what triples are included to describe those resources. That is why, the WHERE block is optional in this query form

Listing 1. SPARQL query to retrieve plants from the linkedplants database

```
1 #Prefix declarations, for abbreviating URIs
2 PREFIX plant: <http://www.linkedplants/plants>
3 #Dataset definition, stating what RDF graph(s) are being queried
4 FROM <http://www.linkedplants.com/data/plants.rdf>
5 #A result clause, identifying what information to return from the query
6 SELECT *
7 #The query pattern, specifying what to query for in the underlying dataset
8 WHERE {?planttype plant:planttype ?name.}
9 #Query modifier, rearranging query results
10 ORDER BY ?name
```

The third Linked Data principle promotes the use of SPARQL, which is useful for querying relationships that are explicitly represented in data. With the respect to geographic information, RDF originally did not support representation of geospatial data and concepts. Therefore, geospatial relationships were implicit and cannot easily be queried.

Geo- semantics and RDF

The first attempt to create a geo extension to RDF was made in 2003, when the W3C Semantic Web Interest Group issued the Basic Geo Vocabulary (W3C Semantic Web Interest Group, 2006) which provided means to represent WGS 84 points in RDF. This work was further extended, and in 2007, the W3C Geospatial Incubator Group (Lieberman et al, 2007) released the GeoOWL ontology, which supported the description of points, lines, rectangles, and polygon geometries and their associated features. However, this ontology also supported only WGS 84 reference system.

The abovementioned ontologies allow representation of geographical data in RDF. Nevertheless, in order to enable spatial reasoning in querying of RDF data, there is a need to include the support of spatial relationships in query language and spatial indexing in triple stores. The release of the GeoSPARQL standard by OGC in 2012 became a breakthrough achievement that brought qualitative and quantitative spatial reasoning to the Web of Data (Battle & Kolas, 2012). This effort used a combination of well-understood and widely used OGC standards such as Geography Markup Language (GML) and well-known text (WKT) literals for representation of geospatial data. The OGC's Simple Features, Egenhofer's 9-intersection model, and RCC8 were used as topological relationship vocabularies and ontologies for qualitative reasoning. So far, several triple stores supporting GeoSPARQL have been implemented including Parliament (<http://parliament.semwebcentral.org/>), Strabon (<http://www.strabon.di.uoa.gr/>) and OpenLink Virtuoso (<http://virtuoso.openlinksw.com/>). These software product provide spatial indexing capabilities which allow computationally efficient access to RDF-based geospatial data.

2.2 Linked Open Data cloud and its geospatial content

Linked Data initiatives promote the adoption of semantic formats. This has enabled users to publish structured data online creating a global data space. In this context, several collaborative projects have emerged, resulting in a growing number of freely available knowledge bases. The W3C SWEO Linking Open Data community project is aimed at publishing various open data sets as RDF and setting RDF links between data items from different data sources. So far, the project has published 570 data sets that are connected by 2909 link sets. Together these resources form the so-called the Linked Open Data cloud, a collection of data sets published under Creative Commons or Talis licenses (LinkingOpenData, 2014). The size of the cloud is large; it contains more than 30 billion triples, which makes it an outstanding source of knowledge.

Ballatore & Bertoloto (2013) have selected and surveyed the five most prominent datasets that have a global scope, are mostly generated through crowdsourcing, released under open licenses, and which are available as fully downloadable dumps in RDF and OWL. The list includes both geographic data resources (e.g. GeoNames and OpenStreetMap), and more general-purpose data sources but containing valuable geographic knowledge (e.g. DBpedia and Freebase).

Selected resources are used as a starting point to explore the Linked Open Data cloud in this thesis. The resources and their description are as follows.

DBpedia is a crowdsourcing collaborative effort mainly led by universities of Leipzig and Berlin, aimed at extraction of information from Wikipedia (Bizer et al., 2009). This is one of the leading projects of the Semantic Web. Altogether the DBpedia 2014 release consists of 3 billion pieces of information (RDF triples) out of which 580 million were extracted from the English edition of Wikipedia, 2.46 billion were extracted from other language editions (Dietzold, 2014). The English version of the DBpedia knowledge base describes 4.58 million things. For each entity, DBpedia defines a globally unique identifier, a URI that can be dereferenced according to the Linked Data principles (Bizer et al., 2009). The content of DBpedia overlaps with various open-license datasets that are already available on the Web. This fosters data publishers to establish RDF links from their data sources to DBpedia, which makes DBpedia a central interlinking hub of the Web of Data.

Geospatial information on DBpedia contains description and locations of 735,000 places (including 478,000 populated places). Geo-coordinates are expressed using the Basic Geo Vocabulary and the GeoRSS Simple encoding of the W3C Geospatial Vocabulary. The former expresses latitude and longitude components as separate facts, which allows for simple areal filtering in SPARQL queries. The data is exposed via a Virtuoso powered SPARQL endpoint.

DBpedia also provides a cross-domain ontology. The ontology was manually created from the most commonly used infobox templates within the English edition of Wikipedia. However, due to the crowdsourcing nature of the original content of Wikipedia, the DBpedia ontology is shallow and lacks some classification hierarchy (Bizer et al., 2009). This drawback hampers semantic interoperability between data entities, the same things can be termed differently depending on an author of content.

GeoNames contains over 10 million toponyms categorized into one out of nine feature classes and further subcategorized into one out of 645 subclasses. This gazetteer integrates geographical data such as names of places in various languages, elevation, and population from numerous data sources. Examples include National Geospatial-Intelligence Agency's (NGA), the U.S. Geological Survey Geographic Names Information System (GNIS), other national mapping and statistics agencies, and crowdsourcing projects (GeoNames., n.d.).

LinkedGeoData (LGD) is an effort that aims to convert into RDF and republish vector data collected by the OpenStreetMap project according to the Linked Data principles (Auer et al., 2007). The knowledge base is maintained and updated on regular basis. Currently the OSM data contains more than 2.5 billion nodes as of October 2014 (OpenStreetMap statistics, 2014). LGD as the biggest RDF-based geospatial data set available on the Web of Data provides spatial dimension for the LOD cloud. LGD is enriched with links to corresponding resources in DBpedia, Geonames, the World Factbook, UMBEL, EuroStat, and YAGO.

WordNet is a lexical database that is widely used as a semantic network and as an ontology (Fellbaum,1998). It was developed by the Cognitive Science Laboratory at Princeton University and became the most successful linguistic resource available online (<http://wordnet.princeton.edu/>). Even though it has limited coverage of geospatial information and lacks of latitude- and longitude coordinates (Buscardi & Rosso, 2008), the ontology provides a high quality, expert-authored conceptualization of geographic concepts (Ballatore & Bertoloto, 2013). This database contains 117659 'synsets' (groups of synonyms) in English. WordNet is characterized by great connectivity with other Linked Data resources.

GeoWordNet emerged as a response to the lack of geospatial information in the WordNet database. This hybrid project includes a thesaurus, a dictionary, a gazetteer, and a semantic network. It was produced as a result of integration of WordNet, GeoNames and the Thesaurus of Geographical Names (Giunchiglia et al., 2010). This knowledge base contains 3.6 million entities, 9.1 million relations between entities, 334 geographic concepts, and 13,000 (English and Italian) alternative entity names, for a total of 53 million RDF triples.

Yet Another Great Ontology (YAGO) is a large knowledge base with high coverage and precision of ontology. It was automatically extracted from Wikipedia and Wordnet (Suchanek et al., 2008). The category system and the infoboxes of Wikipedia were used as sources of facts in YAGO and then this information was combined with taxonomic relations from WordNet. In 2013, YAGO received temporal and spatial extension and was presented as **YAGO2**. In the second version, original content of YAGO was enriched with data from GeoNames.

As a conclusion, at the time, the Linked Data cloud provides an outstanding source of geospatial data. GeoNames and LinkedGeoData together play a central role in this collection of datasets. They contain URIs for geographical objects which makes it possible to refer to particular geographical features in unambiguous way.

By this work, I aim to investigate how to use both geospatial and non-spatial information available in the LOD cloud in order to overcome semantic heterogeneity of user-generated content.

2.3 Integration of data into the LOD cloud

Several works (Goodwin et al., 2008, de Leon et al., 2010, Shvaiko et al., 2012; Tramp et al., 2011) provide an overview of the process of integration of geospatial data with relevant resources in the LOD cloud. In general, this process can be divided into two distinctive and consecutive steps, namely *triplification* and *enrichment* with outgoing links.

The first step in the creation of a new RDF-based dataset is called *triplification* (Faria Cordeiro et al., 2011), or a process of converting raw unstructured data into a set of RDF triples. Triplification requires two decisions to be made. First is to define exactly what information should be encoded into RDF, or triplified. The second decision concerns the choice of vocabulary to use for semantic representation of the data to be triplified. Emergence of the Linked Open Vocabulary (LOV) project (LOV, 2014) has significantly simplified the process of searching for suitable vocabularies. Under the

2.4 Tools for Integration of data into the LOD cloud

The idea of the Semantic Web is not new. For more than a decade different software vendors, opensource communities, and research institutions have been undertaking a task to develop software solutions that would help semantic data integration. Some of the examples have been selected from a considerable number of existing solutions, based on the review of use cases presented in literature and on the Internet. These tools are roughly categorized in to several classes according to their primarily purpose and functionality.

2.4.1 Conversion tools and RDF generators

These tools are developed to facilitate conversion of data from different formats including relational databases into RDF using ontologies. Functionality of both *conversion tools and RDF generators* overlaps to some extent; in general, *RDF generators* provide access to relational databases as virtual, read-only RDF graphs, when *conversion tools* are meant to help in conversion of datasets from different formats to RDF representation with various serialization. These tools are as follows:

The Datalift (Shafre et al., 2012; Dtalift, 2014) is an open software platform able to convert raw structured data coming from various formats (databases, CSV, XML, RDF, RDFa, GML, Shapefile, and others) into semantic data interlinked on the Web of Data. The main functionality of the platform includes selection of ontologies for publishing data, converting data to RDF using the selected ontology, publishing linked data, interlinking data with other data sources, controlling access. The software was developed by a research and development project launched in 2010 and funded by the National Research Agency (ANR). INRIA, National Research Institute on Computer Science and Control is a leading institution in the Datalift project.

TripleGeo (Patroumpas et al., 2014a; TeipleGeo, 2014) is an open-source utility developed by the Institute for the Management of Information Systems at Athena Research Center under the EU/FP7 project GeoKnow: Making the Web an Exploratory for Geospatial Knowledge (<http://geoknow.eu/>). TripleGeo is generic purpose tool that can extract geospatial features from various sources and transform them into triples for subsequent loading into RDF stores. It has wide support of geospatial data representation including GeoSPARQL. It also provides functionality for on the fly geographic reprojection into different Coordinate Reference Systems, and supports a range of standard geographic formats and widely used DBMSs as input.

Triplify (Auer et al., 2009) is a small open-source PHP plugin for Web applications, which converts data from relational databases into RDF and JSON. This software component can be easily integrated into wide-range of Web applications where conversion from RDB to RDF is needed. It is a very lightweight software consisting of less than 500 lines of code. It was developed at University of Leipzig.

D2RQ (Bizer & Seaborne, 2004) offers RDF-based access to the content of relational databases without having to replicate it into an RDF store. The list of capabilities includes querying a non-RDF database using SPARQL, accessing the content of the database as linked data over the Web, creation of custom dumps of the database in RDF formats for loading into an RDF store, accessing information in a non-RDF database using an Apache Jena API.

2.4.2 RDF validators

Validators check RDF datasets for syntax errors, undefined classes or properties, inconsistencies, bogus inverse-functional property values, atypical use of core vocabularies, datatype errors. This type of applications prevents users from malformed input, helps to debug RDF outputs produced by RDF generators and exporters. In general, RDF validators orient users as to what form of data would be

considered "valid" and auto-fill to prompt users for valid input. In spite of RDF validators are very similar to other markup validators, the open world constraints placed on RDF languages make validation difficult and less complete than their counterparts in other data formats (Examples of RDF Validation, 2012). Thus, RDF validation helps more to interpret the data rather than to validate. For instance, some of the validators can also produce a graphical output of the input graph, which is a handy and convenient way to give users some information about the underlying structure of RDF data (Rutledge et al., 2005). Some of the most prominent examples are as follows:

RDF Validator by W3C available at <http://www.w3.org/RDF/Validator/>. This online service allows checking and visualizing of RDF documents. It was developed by Eric Gordon Prud'hommeaux in 2004 (Prud'hommeaux & Lee, 2004)

RDF Alerts is a general purpose Semantic Web/Linked Data validator developed by DERI Galloway. It is available as a web service at <http://swse.deri.org/RDFAlerts/>

Eyeball is a part of Apache Jena framework for checking RDF models (including OWL) for common problems.

Vapour is a linked data validation service (<http://validator.linkeddata.org/vapour.>) to check whether semantic web data is correctly published according to the current best practices, as defined by the Linked Data principles.

2.4.3 Semantic Web browsers

Browsers help to browse and navigate through data published as Linked Data on the Web. This type of applications help to discover data, thus, facilitating it serendipitous re-use (Berners-Lee et al., 2006). Berners-Lee explains this as: "The goal then is that, as with the HTML web, the value is the re-use of information in ways that are unforeseen by the publisher and often unexpected by the consumer" (Berners-Lee et al., 2006). Some of the browser are able to navigate an unbounded set of data sources available on the Web, when others are application- or domain- specific, and therefore their capabilities to browse Linked Data resources are limited (Alahmari et al., 2012). The list below describes only general purpose linked data browsers.

The Tabulator (Lassila, 2006; Berners-Lee et al., 2006; 2007) is an attempt to develop a generic-purpose linked data browser undertaken by Decentralized Information Group Computer from Science and Artificial Intelligence Laboratory at Massachusetts Institute of Technology. It was originally written "as a linked data browser, designed to provide the ability to navigate the web of linked things without any domain-specific programming by the user or the information provider" (Berners-Lee et al., 2006). The Tabulator has some visualization capabilities providing several Views such as Map View, Table View, and Calendar View as different tabs to visualize different types of data.

The Disco - Hyperdata Browser (Bizer & Gauß, 2007) is a simple server-side browser for navigating the Semantic Web as an unbound set of data sources. It does not require installation and can be accessed as Web service. The browser allows navigation between resources by dereferencing HTTP URIs and by following rdfs:seeAlso links rendering all information it finds as HTML. It was developed at the Free University of Berlin, Germany.

OpenLink Data Explorer (ODE) (<http://ode.openlinksw.com/>) is a browser extension (available for Chrome, Firefox, Safari and Opera) allowing users to explore raw data and relationships of a Web page. ODE is also available as a server-side application, which works with any Web browser, as part of the OpenLink AJAX Toolkit.

2.4.4 Triplestores

Databases built for the storage and retrieval of triples using semantic queries are called *triplestores* (Rusher, 2003). There are several types of triplestore implementations, some of the solutions have been developed as native subject-predicate-object database engines from scratch, while others have been built on top of existing commercial relational database engines. The list of existing implementation is quite impressive and includes about 50 different solutions. In this thesis, we focus only on the implementations supporting OWL, spatial indexing and geospatial standards such as OGC's Well Known Text (WKT) or GeoSPARQL. This list is as follows:

Apache Jena (<http://jena.apache.org/>) is an open source Semantic Web framework for Java. It provides an API to extract data from and write to RDF graphs, a SPARQL 1.1 compliant engine built using ARQ, a triplestore, and an ontology API to work with models, RDFS and OWL. Jena has a geospatial extension that enables spatial reasoning in queries. It supports 2 types of RDF representation of geo data, Basic Geo Vocabulary and Well Known Text (WKT). However, Jena provides interface for consuming all kinds of custom geo predicates. A user can simply add predicates to let the software recognize them using EntityDefinition module.

Parliament (Battle & Kolas 2012) uses Jena and modified ARQ query processor, which enables the use of GeoSPARQL. It is an almost complete implementation of GeoSPARQL. It supports both the geo:asWKT and geo:asGML predicates. Development of Parliament was started under the name DAML DB, and was further extended by Raytheon BBN Technologies. It was released under the BSD license in June, 2009. This software uses an innovative data storage scheme, a unique index that allows both fast insertion and fast query, in contrast to most triple stores that favor query speed at the expense of insertion (Kolas et al., 2009).

Strabon is an academic prototype software (Kyzirakos et al., 2012) developed specifically for spatiotemporal RDF data (Patroumpas et al., 2014b). It uses stRDF and stSPARQL, spatially extended versions of RDF and SPARQL, which have been developed independently from GeoSPARQL. Nevertheless, Strabon supports the OGC standards for WKT and GML literals, which makes it partially compliant to GeoSPARQL standard. In addition, Strabon is built on PostgreSQL/PostGIS technology, which enables spatial aggregate functions and triple update commands missing in GeoSPARQL (Patroumpas et al., 2014b). The development of Strabon started in the context of European FP7 project SemsorGrid4Env (Semantic Sensor Grids for Rapid Application Development for Environmental Management). Starting September 2011, Strabon is being utilized and extended with new functionalities in the FP7 project TELEIOS (Virtual Observatory Infrastructure for Earth Observation Data).

uSeekM (Patroumpas et al., 2014b) is an extension library designed for triple stores that offer a Sesame (<http://rdf4j.org/>) compatible API. Wrappers provide most of its functionality. One of them is called IndexingSail and extends an RDF database with indexing and querying capabilities, adding efficient geospatial support, text search, and resource based search. The module indexes spatial data in an R-Tree, Quadtree or Geohash based index, and text data in an inverted-index. uSeekM also requires PostGIS extension to built GiST spatial index. These two indexing schemes help uSeekM to support all OGC geometry types and most operations in the GeoSPARQL standard.

2.4.5 Discovery and establishing of semantic links between data sets

As it was already explained, interlinking of data sets helps to discover more things on the Web of Data. The tools described above provide functionality for data browsing, conversion and storage, but definitely lack means that would support data publishers in setting RDF links to other data sources on the Web. **Silk - A Link Discovery Framework for the Web of Data** is a tool for discovering relationships

between data items within different Linked Data sources (Volz et al., 2009). It uses the declarative Silk - Link Specification Language (Silk-LSL) to allow data publishers specifying what types of links should be discovered between data sources, and what conditions data entities must meet to be interlinked. It provides a number of methods such as ontology mapping and Semantic Similarity Measurement that can be applied to multiple properties of an entity or related entities. SILK is an open source project powered by Assembla. SILK Workbench is the main component of the framework, which is available as a web application.

2.4.6 Query builders and constructors for SPARQL

This type of applications provides visual interfaces that guide and assist users in the process of query construction. Some of the examples are **iSPARQL** (<https://github.com/openlink/iSPARQL>), **SPARQL VIEWS** (https://www.drupal.org/project/sparql_views; Clark, 2010), **SparqlFilterFlow** (<http://sparql.visualdataweb.org/>). All of them offer quite similar functionality such as drag-and-drop visual query building, storing of previous queries, a set of useful predefined queries and query patterns. In addition, Apache Jena framework offers **SPARQLer**, a general-purpose query processor able to validate SPARQL queries. It is a helpful tool to check the correctness of a constructed SPARQL query before it is run against a SPARQL endpoint.

This section shows several software products that have been developed to provide functionality to support users at every stage of data integration workflow. RDF generators and validators help in the process of data triplication. Once data is converted into RDF it can be stored in purpose built databases, namely triple stores. Information stored in triplestores is queried through SPARQL endpoints. Different query builders and constructors provide useful user interfaces facilitating the creation of query statements. Retrieved results can be visualized in a number of RDF data browsers. Methods and functionality needed for semantic enrichment and reconciliation of datasets can be found in link discovery software such as **Silk - A Link Discovery Framework**.

The first research sub questions directly addresses available software products and tools. Further, in this work several tools will be evaluated in terms of applicability for the construction of the proof of concept.

2.5 Visualization of RDF data and SPARQL result set

Humans understand visual representations of data notably faster and more effectively in comparison with doing so by reading the textual representation of the same data (Deligiannidis et al., 2007). Visualization tools are meant to present, transform, and convert data into a visual representation that is approachable by users. However, Linked Data by definition exhibits a graph structure, which is inherently complex to evaluate and interpret (Sobol et al., 2014).

Several software tools were identified from the literature (Graves, 2013; Lemmens & Kessler, 2014, Dadzie & Rowe, 2011) and on the Web. The list below is not meant to present all existing tools, instead it includes software products that don't require installation, available for free of charge, and provide functionality for dynamic visualization of RDF and SPARQL data.

RDF Gravity (Goyal & Westenthaler, 2009; <http://semweb.salzburgresearch.at/apps/rdf-gravity/index.html>) is a software tool built for representation of RDF graphs at Salzburg Research Forschungsgesellschaft mbH. Its functionality allows visualizing RDF and OWL ontologies as well as RDF-based data. In addition, this software makes it possible to visualize multiple RDF datasets allowing filtering of the data to be presented on the screen. Text-based search implemented in the user interface helps to navigate through concepts and instances of an RDF dataset.

Welkin (Mazzocchi & Cicarese, 2008) is a graph-based RDF visualizer developed by Massachusetts Institute of Technology. This is a cross-platform tool, which can run on Windows, Linux and MacOSX. It supports both XML and Turtle/N3 RDF syntaxes providing functionality to turn on/off and modify link strength for individual predicates. With Welkin, a user is able to color code resources and to select and filter nodes directly on their graph-theoretical properties (indegree, outdegree and clustering coefficient).

D3 – Data Driven Documents is a JavaScript library that helps to embed interactive and animated visualization of data into web page elements (Bostock et al., 2011). This library uses HTML, SVG, and CSS standards to manipulate of a web page based on the content of input data. The primary functionality allows visualization of data in the form of graphs, tables, numerous charts, treemaps, and on the map. This library is well documented and it is very easy to find online support and useful examples on the Internet.

D3SPARQL is an open source JavaScript library (project page <https://github.com/ktym/d3sparql>). It performs a SPARQL query using AJAX call, and then transforms the result into JSON and visualizes data with the help of D3 library.

Sgvizler is another JavaScript application visualization of SPARQL results sets (Skjæveland, 2012). This wrapper can be integrated into a web pages providing means to embed SPARQL SELECT queries directly into designated HTML elements. The returned data transformed and visualized with a vast number of charts. In addition, visualisation capabilities can be extended with JavaScript visualization tool-kits, for example D3.

By way of conclusion, RDF data is a graph data; therefore, the main functionality of the tools is to visualize RDF datasets in the form of graphs. In the case of a SPARQL result set, the returned data by default comes in XML-based format with a schema which is not common for non-semantic web applications (Lemmens & Kessler, 2014). Therefore, the data should be first transformed into JSON format and then it can be visualized with the help of a number of visualization tools.

Chapter 3. Information in Emergency Management

This section starts with explanation on how and what Inter-Agency organizations facilitate joined efforts of humanitarian agencies. Then, the second subsection overviews the role of users from all over the Web who help to coordinate and combine mass efforts of online swarm to aid relief operations. In the last subsection, different approaches towards harnessing Linked Data for the purpose of Emergency Management are given.

3.1 Information management in Emergency Management

Timely, relevant and reliable information is vital for the success of an emergency response. The need for a coordinated information management among all the parties involved in emergency relief operations is widely recognized by the humanitarian organizations (Van de Walle et al., 2009; McDonald & Gordon, 2008; Larsen, 2007).

In a nutshell, information management (IM) deals with a range of activities aimed at collection, processing, analyzing and disseminating of the result among stakeholders. The effectiveness of the response directly depends on the ability of the humanitarian community to collect, analyze, disseminate and act on key information. However, a huge number of stakeholders involved in emergency relief actions complicates a problem of effective information management. For instance, the number of Non-Governmental Organizations – mostly relief and development groups – working in Haiti after the earthquake of 12 January 2010 reached 10,000 organizations (Bradly, 2012). This figure did not include the affected population, governmental, intergovernmental, and international humanitarian agencies. Each of these categories, in turn, is not a coherent entity, but a collection of individuals, each of whom uses different information to address different goals in a unique context. Therefore, organization of coordinated information management is a very difficult and important task. Several efforts have been made to organize available information relevant to a disaster.

The IASC (Inter-Agency Standing Committee) established in 1992, provides a broad representation of operational international humanitarian agencies consisting of four main “branches”: The United Nations, International Organization for Migration, the Red Cross Red Crescent Movement and NGOs. In December 2005, the IASC promoted so called “cluster approach” (Inter-Agency Standing Committee, 2006; McDonald & Gordon, 2008; Walle & Dugdale, 2012) to the organization of emergency response operations. This approach implements the division of responsibilities among participating organizations into clusters. Figure 3 illustrates these clusters and corresponding responsible organizations. As can be seen from the Figure, information management plays a key role in the task of coordination of efforts among clusters. The UN’s Office for the Coordination of Humanitarian Affairs (OCHA) facilitates this coordination.

The UNOCHA is an inter-agency body established in 1991 and responsible for the work with operational relief agencies to ensure that there are no gaps in the response and that duplication of effort is avoided (UNOCHA, 2006). The key information that is important to assess and ensure that humanitarian needs are met in any emergency/disaster are: 1) to know which organizations (Who), 2) are carrying out what activities (What), 3) in which locations (Where). This approach refers to as 3W (Who does What Where). The Who does What Where information (3W) is one of the key elements and core products necessary during a disaster response. UNOCHA has developed humanitarian information systems such as ReliefWeb, the Integrated Regional Information Networks (IRIN), Information Management Units (IMUs) and Humanitarian Information Centers (HICs) (Van de Walle et al., 2009). The main functionality of these services ranges from the gathering and collection of information and data, to its integration, analysis, synthesis, and dissemination via the Internet and

other means. All of these services are recognized as essential in the coordination of emergency response among partners in the humanitarian community.

ReliefWeb (<http://www.reliefweb.int>) is the main online gateway to information on humanitarian emergencies and disasters. Using ReliefWeb, OCHA provides practitioners with information on complex emergencies and natural disasters worldwide from more than thousand sources, including UN, governments, NGOs, the academic institutions, and the media. ReliefWeb combines final reports, documents, and reports from humanitarian partners, providing a global repository one-entry point for emergency response information. IRINs gather information from a range of humanitarian and other sources, providing context and reporting on emergencies and at-risk countries. IMUs and HICs collect, manage, and disseminate operational data and information at the field level, providing geographic information products and a range of operations databases and related content to decision makers in the field as well as headquarters.

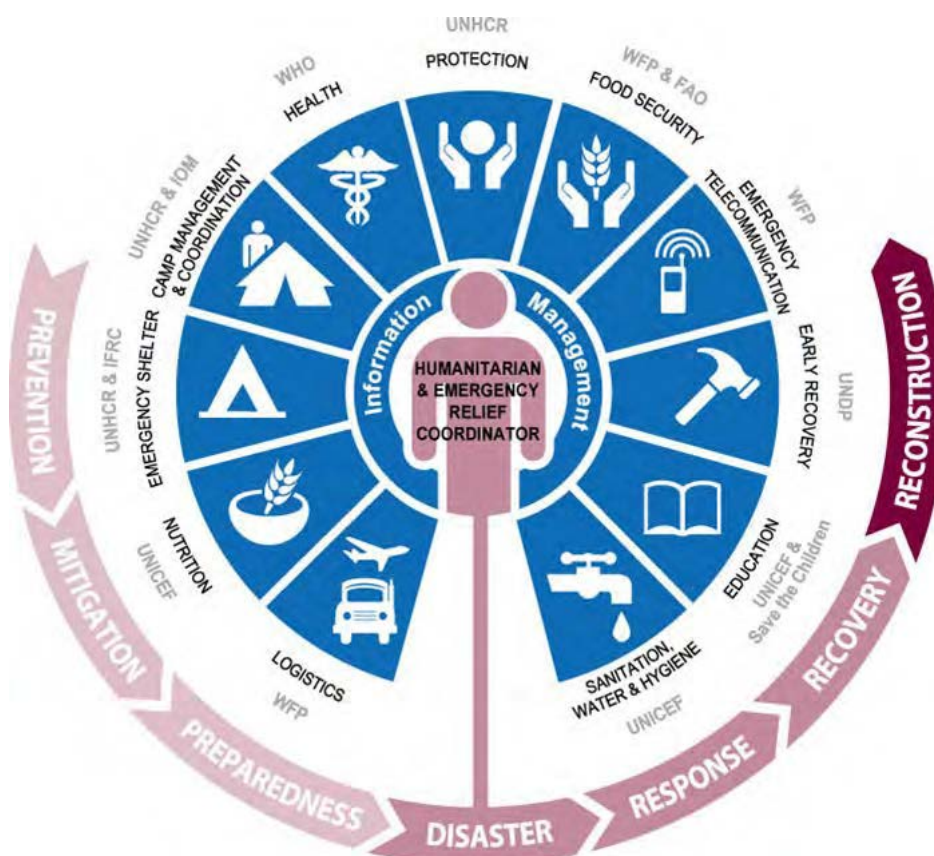


Figure 3. Information Management at the core of Humanitarian Decision Making Process for all the Clusters throughout Crisis/Disaster Management phases (Credit: UNOCHA)

3.2 User Generated Content and crowdsourcing in Emergency Management

The emergence of so-called Social Web or the Web 2.0 was due to a nexus of inventions that enabled increased user involvement in the process of online content generation. Collaborative tools provided by web platforms, such as numerous Wikis, OpenStreetMap project and many others, allowed coordinating and combining of mass efforts made by an online swarm. The term crowdsourcing, coined by Jeff Howe in a 2006 Wired article (Howe, 2006), refers to the use of a group of people on the Web to solve a problem, to develop an idea or just to share, report and communicate a story. With the increasing adoption of smartphones with multiple sensors and a constant Internet connection (Burke et al, 2006), humans as soft sensors became valuable sources of information in the context of many crowdsourcing initiatives.

Crowdsourced relief

The concept of crowdsourcing was adopted in many domains and emergency management was not an exemption (Boulos et al., 2011). In general, experts agreed on the fact that Haiti 2010 earthquake was the first disaster where large scale crowdsourced information was collected (Zook et al., 2010). Mission 4636 (Munro, 2013), a real-time humanitarian crowdsourcing initiative, processed 80 000 text messages (SMS) sent from within Haiti following the 2010 earthquake. This project utilized functionality of two online platforms, HaitianQuake and Ushahidi. HaitianQuake helped to collect information on missing and found people. Ushahidi allowed requesting help via several media channels. Information from both sources was combined and mapped using Ushahidi mapping interface.

Ushahidi, which means “witness” in Swahili, was initially released as a Google Maps mash-up to map reports of violence after the Kenyan post-election fallout in January 2008 (Poblet et al., 2014; Okolloh, 2008). The idea behind the website was to harness the benefits of crowdsourcing information and facilitate the sharing of information in an environment where rumors and uncertainty were dominant. This distributed system provides means for collection of help requests issued by disaster-affected population. People in need can report their concerns via SMS, e-mail, mobile application and the website. A swarm of online volunteers manually verifies the collected requests. The content of each of the messages received by the platform is classified based on the system of ten classes covering the most common types of people’s needs, for example, a water shortage or a need for medical equipment. In addition, every message is geotagged, which makes it possible to put them on the map.

Crowdsourced mapping

Another type of crowdsourcing initiatives that take place during a disaster response is aimed at updating of maps. This was also the case in Haiti (Hattotuwa & Stauffacher, 2011). CrisisCommons, a volunteer driven web-based community, within days created the most comprehensive and up-to-date maps of the country using OpenStreetMap, an online collaborative mapping platform. Thousands of volunteers from around the world contributed to the rapid creation of maps, using different sources of information including situation reports, proprietary databases and satellite imagery. The lessons learned from the Haiti experience of crowdsourced mapping for the purpose of emergency management led to creation of The Standby Task Force (SBTF), an organization that brings digital volunteers together, forming a flexible, trained and prepared network to deploy in crises (About Standby Task Force, 2014).

Summing up, crowdsourcing for the purpose of emergency management comes in several flavors. First, the power of the crowd is used to gain knowledge about needs of people directly from the affected population. Second, online volunteers help humanitarian staff to process and combine information received from disaster-struck area, thus facilitating information management tasks. Finally, the updating of base maps is performed by using collaborative mapping tools such as OpenStreetMap.

3.3 Linked Data for Emergency Management

As it is shown in *Section 3.1*, information management during a catastrophic event deals with the uneasy tasks of helping out various humanitarian agencies to communicate their work, share information and facilitate synchronization of efforts between clusters. These tasks require integration of heterogeneous information that comes in different formats from different clusters. Moreover, recently emerged technologies have given tools and methods to harness the power of the crowd for the purpose of emergency management. *Section 3.2* explains that crowdsourcing initiatives produce valuable information about the needs of people in an affected area and help to increase situational

awareness via collaborative mapping. Since Linked Data enables easy data manipulation and loose integration of heterogeneous information, it makes it an excellent candidate for bringing data generated by volunteers and humanitarian organizations to one global space of interoperable data.

The idea of Linked Data is almost 10 years old. During this time, several works (Mijovic et al., 2013; Ortmann et al., 2011; Terpstra et al., 2012; Schulz et al., 2012; Borges et al., 2011) have undertaken the task to bring semantic technologies into Emergency Management information systems. Based on the analysis of those works several kinds of solutions can be identified.

Vocabularies to annotate EM information

Ortmann et al., 2011 describe an EM information system where data gathered through a conventional EM is transformed into RDF and combined with RDF-based representation of Ushahidi data. The authors have developed a Management of a Crisis (MOAC) vocabulary that helps integration between traditional data sources and user generated content. The vocabulary describes a set of classes and properties covering the main EM concepts and notions. In its core, the MOAC is a lightweight vocabulary; it has just 70 classes and 30 properties to describe EM related information. These properties and types are loosely grouped in three sections without any sophisticated hierarchical structure. Emergency Management Section covers related notions to describe complex emergencies and security incidents. Then, Who What Where (3W) Section explains the basic terminologies about how disaster information managers can identify which organizations (Who) are carrying out what activities (What) in which locations (Where). Inter Agency Standing Committee (IASC) Emergency Cluster Approach Section contains terms used to describe emergency humanitarian clusters like Shelter Cluster or Health Cluster. In addition, it includes classes created specifically for Ushahidi categories used in Haiti.

The work of Keßler & Hendrix (forthcoming) follows the development of the MOAC vocabulary. The authors created the HXL vocabulary—officially entitled Humanitarian eXchange Language (HXL) Situation and Response – a project by UNOCHA aimed at refining data management and exchange for disaster response. The HXL focuses on rather quantitative information than qualitative. The intention of the authors is to allow valuable numerical data to be used directly to generate reports, maps, and interactive dashboards. It consists of five sections: geolocation section, humanitarian profile section, metadata section, response section and situation section. In general, this vocabulary is more structured than MOAC, however it follows the same logic of Who What Where. The major difference between the MOAC and the HXL is in the absence of classes corresponding to Ushahidi categories in the HXL vocabulary.

The MOAC and HXL are purpose built vocabularies created to satisfy specific needs of EM data integration. More general vocabularies such as Dublin Core and SIOC are meant to describe very common characteristics of information, for instance, an author, date of creation, coverage etc.

The Dublin Core community, started as a series of workshops in Dublin, Ohio in 1995, brought together librarians, digital library researchers, content experts, and text-markup experts “to promote better discovery standards for electronic resources” (The Dublin Core Metadata Initiative, 2014). This effort led to creation of the Dublin Core Metadata Element Set consisting of 15 basic terms. This vocabulary was published as IETF RFC 5013, ANSI/NISO Standard Z39.85-2007, and ISO Standard 15836:2009. The Dublin Core Metadata Initiative Metadata Terms is an extended vocabulary built on the Dublin Core Metadata Element Set, which includes additional properties and classes.

If Dublin Core was born to facilitate interoperable metadata exchange for librarians, the SIOC vocabulary emerged to help online communities in linking information about their structure and contents. SIOC stands for Semantically-Interlinked Online Communities, a project, which was started

in 2004 by John Breslin and Uldis Bojars at DERI, NUI Galway (Breslin et al., 2009). The organization became a member of W3C in 2007. One of the main products of this project was The SIOC Core Ontology. This ontology made it possible to describe the main concepts and properties of information from different online communities on the Semantic Web, for instance, webblogs, message and discussion boards, social networks etc. With SIOC a user can model facts such as *"Here is an item written by Alice that has been commented on by Bob at <http://example.org/aliceblog>"* and *"Alice is the moderator of <http://example.org/aliceboard> while Bob is a simple reader"*, using the same model wherever the data comes from (Breslin et al., 2009).

Integration of EM information with LOD

Comprehensive solutions that combine crowdsourced and official data sources and link them to relevant entities in the LOD cloud are described in Schulz et al., 2012 and Borges et al., 2011. This approach allows semantic enrichment of information through the semantics captured in ontologies and taxonomies of the LOD resources. Moreover, the LOD cloud as an unbounded data repository allows storing and integration of past and newly generated data.

Terpstra et al., 2012 propose a system where unstructured information from social media such as Twitter messages is converted into RDF and then linked. Conversion of messages into RDF implies the use of natural language processing (NLP) and annotating. Mijovic et al. (2013) test the capability of DBpedia Spotlight, a tool for automatically annotating mentions of DBpedia resources in text. The authors conclude that the more specific the words are in the text the better results are produced. This is because of so-called "common words" problem, a tendency of annotation engine to generate very general links to most common DBpedia pages. Therefore, it should be used carefully or only with prior training.

Nevertheless, despite of the great potential of Linked Data, it has its flaws as any other technology. Two main issues concern the quality of data. On the one hand integrating of VGI raises the trustworthiness issue because information collected by a layperson lacks quality assurance processes integral to the data collection. On the other hand, the data presented in the LOD cloud is collected from community driven sources (government portals or other Web resources). Consistency and completeness of the LOD cloud content vary depending on population density; areas with higher population density receive more attention from users than less populated territories (Graham et al., 2014). Another issue is shown in the work of Milis & van de Walle (2007), who notice that new technologies and IT solutions find their way into emergency management only if there is "a member in the crisis management with an IT background" (Milis & van de Walle, 2007). However, the members of emergency management organizations are often non-IT experts (Babitski et al., 2011). This creates additional obstacles for using Linked Data solutions.

Chapter 4. Prototyping the case study

This chapter explains the data and approaches taken to construct a proof of concept for the case study. The case study takes into consideration the VGI generated in the wake of the 2010 Chile earthquake. The aim of the case study is to investigate to what extent this VGI can be enriched with additional semantics coming from LOD and how this enrichment influences on answering queries against this data.

The section starts with a description of the dataset used in the research, providing details on data structure and provenance. The work adopts a heuristic approach towards the formulated research questions. In this approach, the research sub questions are answered based on the experience gained in the development of the proof of concept. In order to bring more formalism in the chapter, the entire work has been broken down into two work packages, each of which is explained separately. However, it is worth mentioning that work packages together form a seamless workflow.

At the end of the work, the answer to the main research question is derived from the synthesis of information acquired from the literature review and the knowledge gained from the experiments with software and tools.

4.1 Data

Data collected by the Ushahidi project during the Chilean earthquake of 2010 has been selected as an example of a disaster related VGI.

The 2010 Chile earthquake ranks as the sixth largest earthquake ever to be recorded by a seismograph. The earthquake took place off the coast of central Chile on Saturday, 27 February 2010, at 03:34 local time (06:34 UTC) (USGS, 2010). It had a magnitude of 8.8 on the moment magnitude scale; the shaking lasted for about three minutes. The disaster mainly affected six Chilean regions (from Valparaíso in the north to Araucanía in the south), that together make up approximately 80% of the country's population. The earthquake also triggered a tsunami, which struck coastal regions in about 30 minutes after the first shock. Talcahuano port was seriously damaged when some coastal towns were completely devastated. The blackout caused by the disaster affected 93 percent of the Chilean population and went on for several days in a number of locations. Official sources reported 525 people lost their lives, 25 people went missing and about 9% of the population in the affected regions lost their homes (BBC, 2010; USGS, 2010).

Content and acquisition

The dataset was composed of 1228 reports which were gathered from the affected population as well as from official sources such as UNOCHA reports, officials' statements and media. This number included only those reports that were approved by the Ushahidi staff. Volunteers checked the credibility of the reports where it was possible. For example, if a report was received from the website or SMS, Ushahidi activists called back or emailed to the reporter. In the case of anonymous reports, incidents were counter-checked by comparing with other sources, for instance, mainstream media. If information appeared credible but could not be verified, it was posted with notes that it was not verified (Okolloh, 2008). In addition, each of the reports was assigned accordingly to a specific category such as buildings collapsed, bridges closed, people trapped, etc. The list of categories was created rather *ad hoc* when a disaster had stricken. Categorization system varied from one Ushahidi deployment to another. Nevertheless, in general, these categories described quite similar things but in a different order and with different degree of details.

The data has been accessed via the Ushahidi API at <http://chile.ushahidi.com/api>, using Postman – Rest client, an extension to Google Chrome Web browser. By default, the Ushahidi server returns data in JSON format. The data has been transformed into a more convenient tabular form using functionality of Microsoft Excel. Example reports (in JSON and tabular) from the original dataset can be seen in Appendix A.

Information about a report comprises a row in a table with 10 columns. Every column describes one attribute of a report. Table 1 explains the content of the columns, corresponding data types and gives example records. Each of the reports has a unique serial number and an incident title presenting a short description of a report. The time stamp of a report can be seen in the column “Incident date”. The field “location” describes the location of a report. Coordinates of this location can be found in columns “Latitude” and “Longitude”. The original message is given in the column “Description”. The field “Category” provides a category assigned to a report. Fields “Approved” and “Verified” give information about the creditability of a report.

Table 1. Structure of a report

Column name	Column description	Data type	Example records
Serial number	Identification number of a report	integer	4389, 3054, 732
Incident title	Short title of an incident described in a report	character	Destroyed building in Talca Arauco falta de agua/ Arauco without water
Incident date	A date when an incident was witnessed	date	5/4/2010 13:22
Location	Description of a location where an incident took place	character	Las Tranqueras, Temuco Chile
Description	Description of an incident/original message	character	El puente itata sobre el rio itata de la localidad de Coelemu, 8va region, se encuentra en mal estado,y no se sabe quien lo va a arreglar! Las autoridade
Category	Shows what a category a report belongs to	character	3.Catastro, 3a. Desabastecimiento de Agua,
Latitude	Latitude of a place described in column location	double	-36.78691
Longitude	Longitude of a place described in column location	double	-73.11358
Approved	State of approval.	Boolean	YES/NO
Verified	Indicated whether a report is verified	Boolean	YES/NO

4.2 Work package 1. Conversion of data into RDF with links to LinkedGeoData

Workflow

The first part of the work has covered two issues: a conversion of a dataset generated by Ushahidi from its original form into RDF-based representation, followed by establishing of outgoing links to relevant resources in the LOD cloud. Figure 4 shows the workflow diagram of this work package. The first steps in the diagram dealt with accessing of data from Ushahidi, which has been described in the previous section. Once the data has been obtained and transformed into tabular form, it can be converted into RDF. In order to do this, two important design decisions have been made.

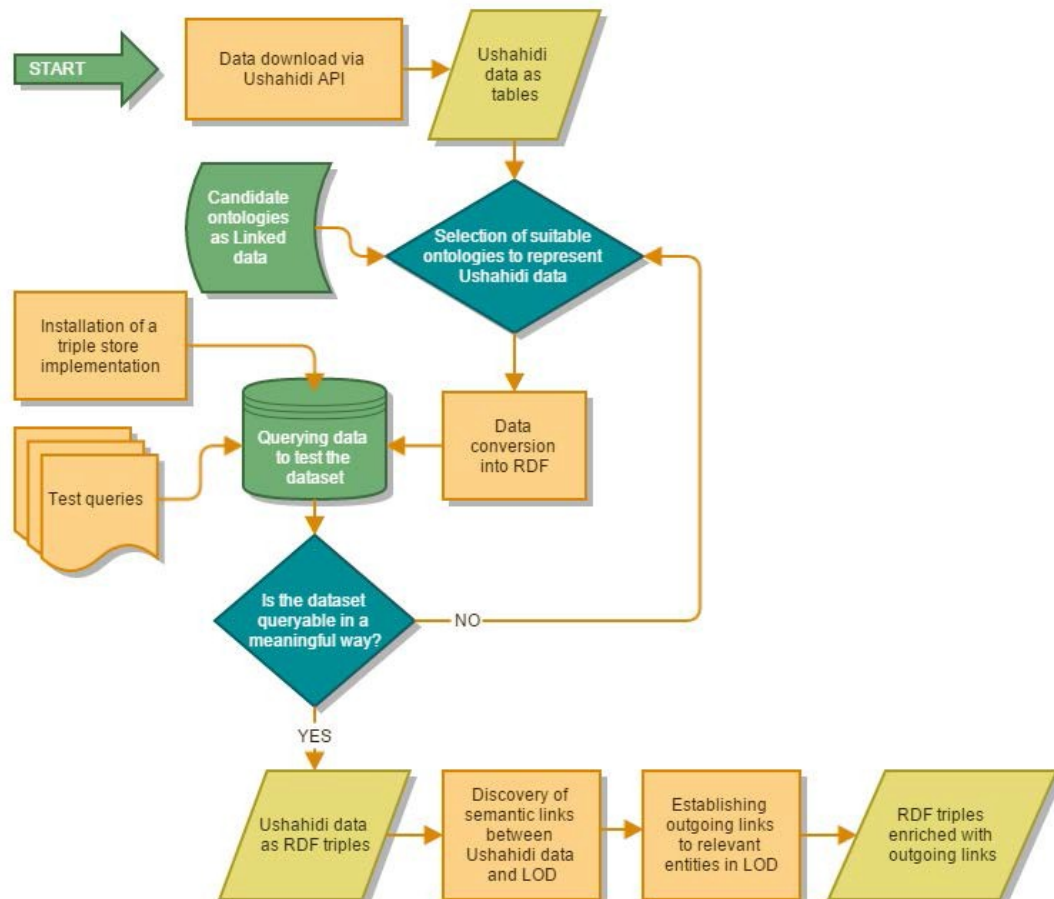


Figure 4. Workflow diagram for the first work package.

The first decision is concerned with the first principle of Linked Data, namely - to use URIs as names for things. In the case of Ushahidi, there were not any agreed upon namespaces and ready-made URIs. Therefore, new URIs have been coined based on serial numbers of the reports. Since the proof of concept has been built in a local computational environment, <http://localhost/chile/> is used as a namespace for the URIs. As a result, triplified reports had a structure where a report URI was always the subject of all the triples describing the report. Corresponding values in columns became objects of those triples. Objects were connected to the subject with semantically explicit predicates.

The second important decision has been made about ontologies and vocabularies used for predicates and objects. From Table 1 it is clear that all the information presented in a report could be converted into RDF as literal objects whether plain or typed.

However, representation of information only in the form of literals makes it difficult to discover semantic links to resources in the LOD cloud. In this case, establishing of semantic links is based only on the matching of strings and substrings. For the sake of further semantic enrichment, it is better to use terms from controlled vocabularies as objects where this is possible. Such an approach allows filtering of data as well as searching for relevant semantic links using ontology matching.

Conversion into RDF

Management of a Crisis (MOAC) vocabulary has been used to encode Ushahidi categories into RDF. In order to identify the MOAC terms corresponding to the categories used for Chilean earthquake an ontology mapping has been performed. Appendix B shows the correspondence between the categories used in Chile and the MOAC terms. If a proper term was absent in the MOAC vocabulary, a substitute term was taken from one of other vocabularies available on the Web. Based on this mapping, every report has been marked with a relevant term.

The Dublin Core vocabulary has been used for predicates to wrap into RDF the content of incident titles, incident dates and locations. The SIOC (Semantically-Interlinked Online Communities) Core Ontology provides a predicate for incident description. Latitude and longitude values are presented using WGS84 Geo Positioning RDF vocabulary. Table 2 summarizes predicates utilized for the description of a report.

Table 2. Table of predicates

Predicate	Attribute of a report
http://purl.org/dc/elements/1.1/title	Incident Title
http://purl.org/dc/elements/1.1/date	Incident Date
http://purl.org/dc/elements/1.1/coverage	Location
http://rdfs.org/sioc/ns#content	Description
http://observedchange.com/moac/ns#subjectlabel	Category
http://www.w3.org/2003/01/geo/wgs84_pos#lat	Latitude
http://www.w3.org/2003/01/geo/wgs84_pos#long	Longitude
http://observedchange.com/moac/ns#approved	Approved
http://observedchange.com/moac/ns#verified	Verified
http://purl.org/dc/elements/1.1/subject	MOAC term for a category

The conversion from tabular form into RDF is performed in OpenRefine. As a result of this operation 17824 triples were generated and exported from OpenRefine as XML/RDF. Table 3 and Listing 2 provide an example of an original report and its RDF-based representation.

Table 3. Report 4349

Column name	Value
Serial number	4349
Incident title	SERVICIO DE SALUD CONCEPCIÓN FUNCIONANDO
Incident date	3/4/2010 11:08:00 PM
Location	Concepcion, Chile
Description	Hospital Guillermo Grant Benavente , Hospital Traumatológico , Hospital de Lota , Hospital de Coronel
Category	4a. Servicios de Salud,
Latitude	-36.8148
Longitude	-73.0293
Approved	YES
Verified	NO

Listing 2. RDF-based representation of report 4349

```

1 <?xml version="1.0" encoding="utf-8" ?>
2 <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
3     xmlns:DC="http://purl.org/dc/elements/1.1/"
4     xmlns:sioc="http://rdfs.org/sioc/ns#"
5     xmlns:MOAC="http://observedchange.com/moac/ns#"
6     xmlns:geo="http://www.w3.org/2003/01/geo/wgs84_pos#"
7     xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
8     xmlns:tisc="http://observedchange.com/tisc/ns/#"
9     xmlns:dcterms="http://purl.org/dc/terms/"
10
11 <rdf:Description rdf:about="http://localhost/chile/4349">
12   <rdf:type rdf:resource="http://observedchange.com/moac/ns/UshahidiReport"/>
13   <DC:title>SERVICIO DE SALUD CONCEPCIÓN FUNCIONANDO</DC:title>
14   <DC:date rdf:datatype="http://www.w3.org/2001/XMLSchema#dateTime">Thu Mar 04 23:08:00 CET 2010</DC:date>
15   <DC:coverage>Concepcion, Chile</DC:coverage>
16   <sioc:content>Hospital Guillermo Grant Benavente, Hospital Traumatológico, Hospital de Lota, Hospital de Coronel</sioc:content>
17   <MOAC:subjectlabel>4a. Servicios de Salud</MOAC:subjectlabel>
18   <geo:lat rdf:datatype="http://www.w3.org/2001/XMLSchema#float">-36.814815</geo:lat>
19   <geo:long rdf:datatype="http://www.w3.org/2001/XMLSchema#float">-73.029257</geo:long>
20   <MOAC:approved rdf:datatype="http://www.w3.org/2001/XMLSchema#boolean">>true</MOAC:approved>
21   <MOAC:verified rdf:datatype="http://www.w3.org/2001/XMLSchema#boolean">>false</MOAC:verified>
22   <DC:subject rdf:resource="http://observedchange.com/moac/ns/#HospitalOperating"/>
23 </rdf:Description>
24
25 </rdf:RDF>

```

Figure 5 illustrates the structure of the report number 4349 before (Figure 5 A) and after (Figure 5 B) the enrichment with the LGD links. In the Figure, green rectangles represent literal values, purple rectangles mean terms from controlled vocabularies and purple triangles illustrate URIs of things. It is clear from the Figure, the initial data model of Ushahidi data allowed only one-to-one cardinality of relationships between data instances (Figure 5 A). In other words, an attribute of a report could possess only one value. Conversion of the reports using RDF created an opening to change the data model to allow more than one description of a place per a report (Figure 5 B). Therefore, after the conversion cardinality of the relationship between a report and its location became one-to-many.

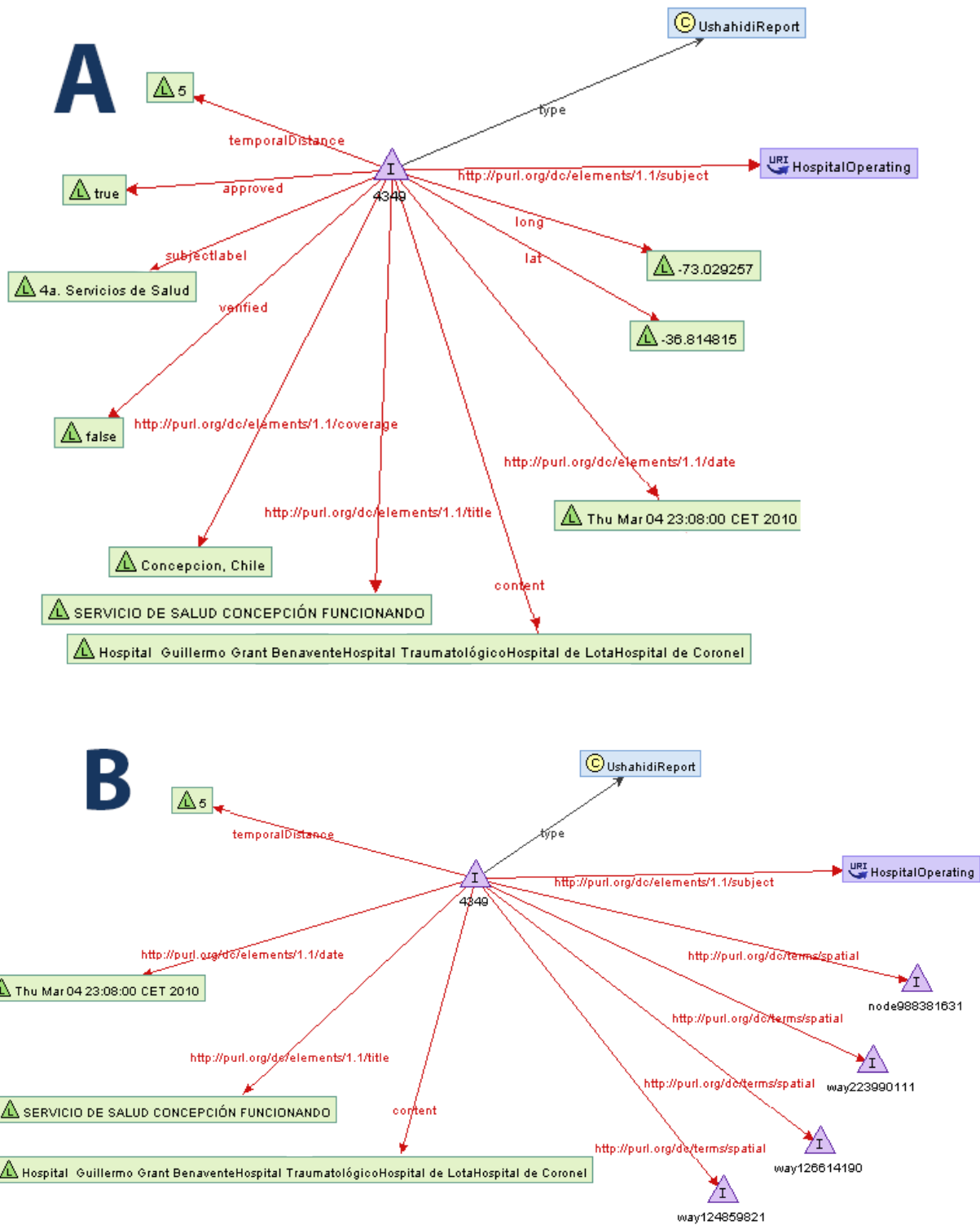


Figure 5. Graph visualization of the report 4349 in the initial state (A) after the conversion (B)

Testing the Data

Once the conversion was done, the triples were uploaded into a triplestore. Then, the data is put to the test whether the dataset is queryable in a meaningful way (see the workflow diagram in Figure 4).

The Parliament triplestore (see *Section 2.4.4*) has been used as a datastore and a SPARQL endpoint. The XML/RDF file has been uploaded using a bulk load function of the triplestore. A set of queries has been run against the local SPARQL endpoint to test integrity of the dataset.

The first testing query selects all the reports describing collapsed structures. This query can be seen in Listing 3. For the sake of convenience namespace URIs are abbreviated with prefixes. In this thesis, examples assume the namespace prefix bindings presented in Appendix E unless otherwise stated.

Listing 3. The first test query

```
1 SELECT DISTINCT ?s
2 WHERE
3 {
4     ?s a MOAC:UshahidiReport.
5     ?s dc11:subject MOAC:CollapsedStructure.
6 }
```

Retrieved data included 64 reports, which was equal to the number of the reports that were categorized as “Estructura Colapsada”. The first five solutions are shown in Figure 6. The overall number of the solutions can be seen at the top of Figure 6 – “count: 64”



Figure 6. Example five reports assigned with a category "Collapsed Structure".

The next query counted the number of categories assigned to a report and selected those reports where the number of assigned categories was greater than two. Listing 4 provides this query. The result was a list of 105 reports that featured more than two categories. The first five results can be seen in Figure 7. Listing 5 illustrates the report number 3970 where 10 categories were assigned.

Listing 4. Selection of the reports with more than 2 categories.

```
1 SELECT DISTINCT ?s (Count(?q) as ?count)
2 WHERE
3 {
4     ?s a MOAC:UshahidiReport.
5     ?s dc11:subject ?q.
6     }
7 GROUP BY ?s
8 HAVING (?count > 2)
```

Count: 105

s	count
http://localhost/chile/4275	"4" ^^<http://www.w3.org/2001/XMLSchema#integer>
http://localhost/chile/4699	"3" ^^<http://www.w3.org/2001/XMLSchema#integer>
http://localhost/chile/4025	"3" ^^<http://www.w3.org/2001/XMLSchema#integer>
http://localhost/chile/3711	"3" ^^<http://www.w3.org/2001/XMLSchema#integer>
http://localhost/chile/3790	"10" ^^<http://www.w3.org/2001/XMLSchema#integer>

Figure 7. Results retrieved by the query in Listing 4.

Listing 5. Report 3790 with 10 assigned categories.

```

22585 <rdf:Description rdf:about="http://localhost/chile/3790">
22586   <rdf:type rdf:resource="http://observedchange.com/moac/ns/UshahidiReport"/>
22587   <DC:title>Red Cross Central Warehouse Now Receiving Supplies</DC:title>
22588   <DC:date rdf:datatype="http://www.w3.org/2001/XMLSchema#dateTime">Sun Feb 28 15:26:00 CET 2010</DC:date>
22589   <DC:coverage>Santiago, Seminario 973, Ñuñoa</DC:coverage>
22590   <sioc:content>Receiving foodstuffs at the Chilean Red Cross Central warehouse</sioc:content>
22591   <MOAC:subjectlabel>3. Catastro, 3a. Desabastecimiento de Agua, 3d. Desabastecimiento de Alimentos,
22592   3e. Desabastecimiento de Medicamentos, 4. Respuesta, 4a. Servicios de Salud, 4d. Distribución de Alimentos,
22593   4e. Saneamiento de Agua, 4f. Recepción de Ayuda, 4i. Distribución de Agua</MOAC:subjectlabel>
22594   <Geo:lat rdf:datatype="http://www.w3.org/2001/XMLSchema#float">-33.45493</Geo:lat>
22595   <Geo:long rdf:datatype="http://www.w3.org/2001/XMLSchema#float">-70.625981</Geo:long>
22596   <MOAC:approved rdf:datatype="http://www.w3.org/2001/XMLSchema#boolean">>true</MOAC:approved>
22597   <MOAC:verified rdf:datatype="http://www.w3.org/2001/XMLSchema#boolean">>false</MOAC:verified>
22598   <DC:subject rdf:resource="http://observedchange.com/moac/ns/InfrastructureDamage"/>
22599   <DC:subject rdf:resource="http://observedchange.com/moac/ns/WaterShortage"/>
22600   <DC:subject rdf:resource="http://observedchange.com/moac/ns/FoodShortage"/>
22601   <DC:subject rdf:resource="http://observedchange.com/moac/ns/MedicalEquipmentAndSupplyNeeds"/>
22602   <DC:subject rdf:resource="http://observedchange.com/moac/ns/ServiceAvailable"/>
22603   <DC:subject rdf:resource="http://observedchange.com/moac/ns/HospitalOperating"/>
22604   <DC:subject rdf:resource="http://observedchange.com/moac/ns/FoodDistributionPoint"/>
22605   <DC:subject rdf:resource="http://observedchange.com/moac/ns/WaterSanitationAndHygienePromotion"/>
22606   <DC:subject rdf:resource="http://observedchange.com/moac/ns/NonfoodAidDistributionPoint"/>
22607   <DC:subject rdf:resource="http://observedchange.com/moac/ns/WaterDistributionPoint"/>
22608   <tisc:temporalDistance rdf:datatype="http://www.w3.org/2001/XMLSchema#int">1</tisc:temporalDistance>
22609   <dcterms:spatial rdf:resource="http://localhost:3333/chile/Santiago%2C+Seminario+973%2C+%C3%91u%C3%B1oa"/>
22610 </rdf:Description>

```

Establishing semantic links to the LOD resources

After the successful testing, the data can be enriched with outgoing links to relevant LOD resources. As it was described in *Section 2.3*, semantic links can be established based on one of the three components of geographic information - a spatial, a temporal, and a thematic (or attribute) component.

The spatial component in the case of the Ushahidi data was presented as the location of a report. This information could be used for generation of links between reports and relevant entities in GeoNames and LinkedGeoData. However, analysis of the original data set revealed a great degree of inconsistency of records in the column "location". In general, this column pointed at the place where an incident occurred. However, 36 reports had geographic coordinates instead of place descriptions in this column. Moreover, many place names had typos and misspellings. For instance, one of the largest cities of Chile, Concepcion, had more than 20 different variants of spelling in more than 100 records where it was mentioned. In the initial state, this column had more than 1100 unique records out of total 1228. This meant that almost every record needed to be checked.

Another issue concerned the quality of georeferencing. The Ushahidi platform allowed visualization of the location of a report as a dot on the map. This was done using coordinates from columns “longitude” and “latitude”. Obviously, georeferencing to a point had its drawbacks. For instance, in the case, when a report described an incident that took place in several locations within one administrative unit, it was usually georeferenced to a centroid of a polygon representing this administrative unit. As a result, such a point appeared in the middle of nowhere and lacked any meaning. Figure 8 illustrates such a situation. For instance, a report with the number 4349 lists four operating hospitals. These hospitals are scattered over several cities of Bio-Bio region, including Coronel, Lota and Concepcion (blue dots in Figure 8). However, the report is referenced to a location in Concepcion (red dot in Figure 8).

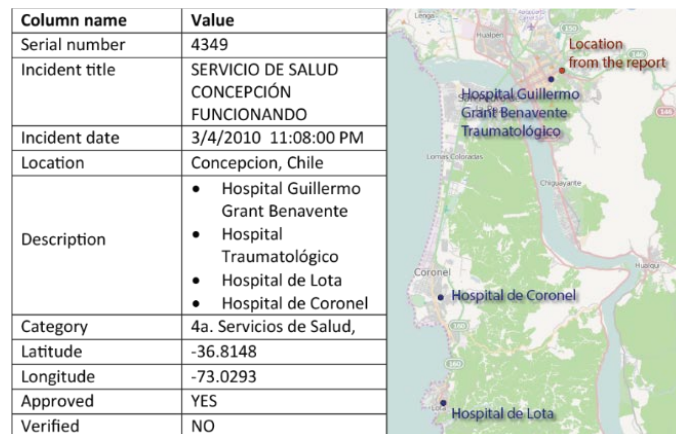


Figure 8. Georeferencing of the report 4349

The first issue, inconsistency of location descriptions, created significant obstacles for automated generation of links. Locations were converted into RDF as literals. This fact limited the available methods for discovery of semantic links to a keyword search only. In the case when data is messy, full of spelling mistakes and typos, it is almost impossible to search for links. The data should first be cleaned and refined. Moreover, the second issue, ambiguous georeferencing, eliminated the possibility to check the locations based on coordinates because coordinates often pointed to centroids of administrative units instead of particular objects. As a result, the amount of manual work required for data cleaning, refining and fixing of all the georeferencing errors presented in 1228 reports was too large for this research.

For the construction of the proof of concept, it was decided to select a number of the reports, which were the most sensitive to the accuracy of spatial descriptions. Information about blocked roads and damaged bridges is of great importance for planning of relief actions since it influences logistic and routing. Information about operating hospitals comprises another group of reports that need careful georeferencing. Location of available social objects such as schools, pharmacies, ATMs and supermarkets is vital for affected population. The idea was to search for spatial objects mentioned in incident descriptions and to enrich manually those reports with outgoing links to corresponding entities in LGD.

Eighty one reports have been deliberately selected and enriched with links to 95 spatial objects mentioned in those reports. The list of the selected reports can be found in Appendix F. The search has been performed using OpenStreetMap search service. OSM uses the same codes for objects as LGD. URIs have common namespace - <http://linkedgedata.org/triplify/>, followed by an object code, for example, <way48592362>. The predicate <<http://purl.org/dc/terms/spatial>> is used to connect the subject (a report URI) and the object (URI of particular spatial object) of a triple describing the

location. The resulting dataset is composed of 773 triples, which have been uploaded into Parliament triplestore.

In addition, due to the fact that the work deals with information collected during a disaster that occurred 4 years ago, it is important to take into consideration temporal information presented in the reports. For the sake of convenience of manipulating with temporal information, it has been decided to enrich reports with information about temporal distance. Temporal distance is an integer value that has been calculated as a temporal difference between actual date of a report and the date when the emergency started. A term “temporalDistance” from spatio-temporal vocabulary (<http://observedchange.com/tisc/ns/#>) has been employed as a predicate for values.

Listing 6 shows the final RDF version of a report with the serial number 4349 (the same report as in Figure 8). It includes all the information from the original report enriched with links to LGD representation of spatial objects described in the report (hospitals) and temporal distance.

Listing 6. Final version of the report 4349.

```

1  <?xml version="1.0" encoding="utf-8" ?>
2  <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
3      xmlns:DC="http://purl.org/dc/elements/1.1/"
4      xmlns:sioc="http://rdfs.org/sioc/ns#"
5      xmlns:MOAC="http://observedchange.com/moac/ns#"
6      xmlns:geo="http://www.w3.org/2003/01/geo/wgs84_pos#"
7      xmlns:tisc="http://observedchange.com/tisc/ns/#"
8      xmlns:dcterms="http://purl.org/dc/terms/">
9
10 <rdf:Description rdf:about="http://localhost:3333/chile/4349">
11   <rdf:type rdf:resource="http://observedchange.com/moac/ns/UshahidiReport"/>
12   <DC:title>SERVICIO DE SALUD CONCEPCIÓN FUNCIONANDO</DC:title>
13   <DC:date rdf:datatype="http://www.w3.org/2001/XMLSchema#dateTime">Thu Mar 04 23:08:00 CET 2010</DC:date>
14   <DC:coverage>Concepcion, Chile</DC:coverage>
15   <sioc:content>Hospital Guillermo Grant Benavente, Hospital Traumatológico, Hospital de Lota, Hospital de Coronel</sioc:content>
16   <MOAC:subjectlabel>4a. Servicios de Salud</MOAC:subjectlabel>
17   <DC:subject rdf:resource="http://observedchange.com/moac/ns/#HospitalOperating"/>
18   <geo:lat rdf:datatype="http://www.w3.org/2001/XMLSchema#float">-36.814815</geo:lat>
19   <geo:long rdf:datatype="http://www.w3.org/2001/XMLSchema#float">-73.029257</geo:long>
20   <tisc:temporalDistance rdf:datatype="http://www.w3.org/2001/XMLSchema#int">5</tisc:temporalDistance>
21   <dcterms:spatial rdf:resource="http://linkedgeoedata.org/triplify/way124859821"/>
22   <dcterms:spatial rdf:resource="http://linkedgeoedata.org/triplify/way223990111"/>
23   <dcterms:spatial rdf:resource="http://linkedgeoedata.org/triplify/way126614190"/>
24   <dcterms:spatial rdf:resource="http://linkedgeoedata.org/triplify/node988381631"/>
25 </rdf:Description>
26
27 </rdf:RDF>

```

4.3 Work Package 2. Construction of queries for a semantic enrichment

The aim of the second work package was to develop and to test a number of flexible and easily customizable SPARQL queries, which could help people to enrich Ushahidi data with additional relevant information from the LOD cloud. Section 3.1 has shown that a number of stakeholders involved in an emergency is considerable, and as a result, it is extremely difficult to foresee information needs of decision makers since they have different goals. Therefore, the main objective of this section is to develop SPARQL queries that would retrieve as much information as possible tracing graph patterns emerged via outgoing links to LinkedGeoData entities.

A workflow diagram of this work package is depicted in Figure 9. As can be seen from the Figure, the query development process has taken place in an iterative manner. The work has started with a construction of an initial query. If a query has generated expected result, retrieved information is added to the triplestore; otherwise, the query is improved following iterative loops.

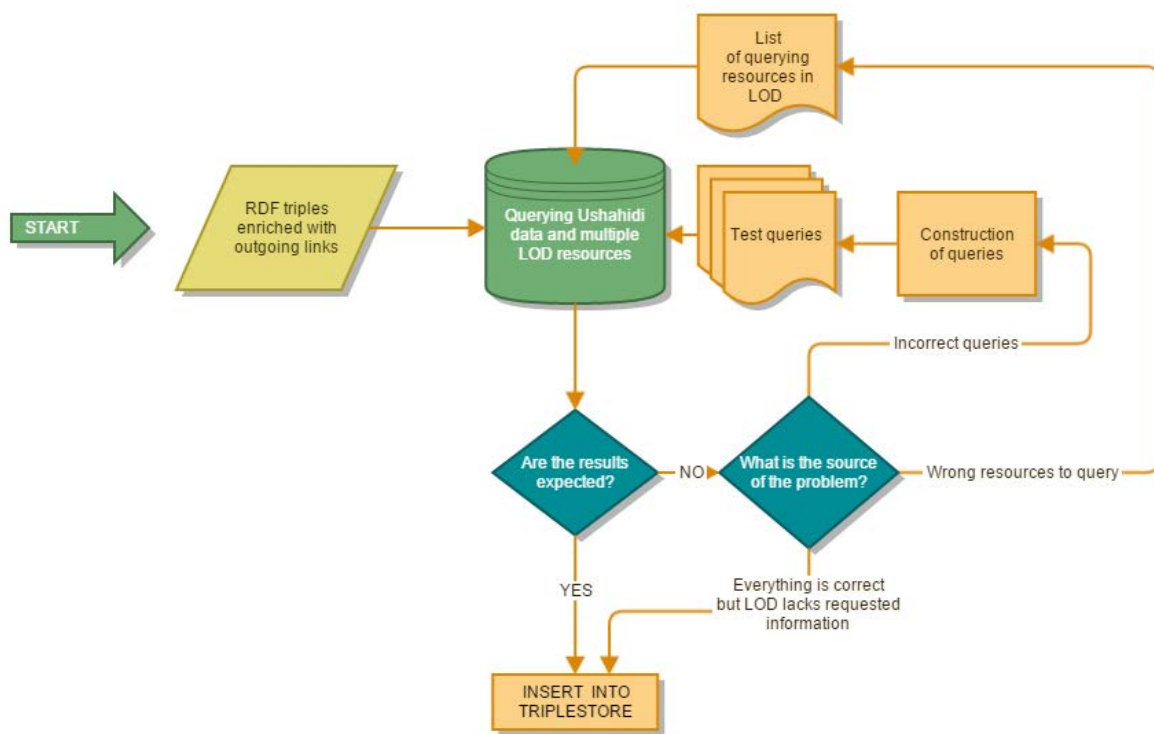


Figure 9. Workflow diagram for the second work package.

Retrieval of LGD data

The first and most obvious use case scenario was to obtain structured information from the LOD cloud following established links to LinkedGeoData. Figure 10 shows all the information available in the LGD database for the object < <http://linkedgeo.org/triplify/way126614190>>. This object is a representation of “Hospital de Coronel” which is mentioned in the report number 4349 (see Figure 8 and Listing 6). It is clear from Figure 10, that the description includes a name, which is Hospital de Coronel, a geometry of the object, and a list of classes the object belongs to. Therefore, this information can be accessed via a SPARQL endpoint of LGD and be used for enrichment of the initial semantics of the report 4349. Consequently, the first SPARQL query of this section selects all the information from LGD relevant to spatial objects and then inserted selected triples into the triplestore.

Hospital de Coronel at LinkedGeoData	
http://linkedgeo.org/triplify/way126614190	
Property	Value
lgdo:changeset	▪ 13919887 (xsd:int)
dcterms:contributor	▪ lgdo:user337684
geom:geometry	▪ lgd-geom:way126614190
rdfs:isDefinedBy	▪ lgd:meta/way126614190
rdfs:label	▪ Hospital de Coronel
dcterms:modified	▪ 2012-11-18T17:02:41 (xsd:dateTime)
rdf:type	▪ spatial:Feature ▪ lgdm:Way ▪ lgdo:Amenity ▪ lgdo:EmergencyThing ▪ lgdo:Hospital
lgdo:version	▪ 2 (xsd:int)

Figure 10. Representation of Hospital de Coronel in LinkedGeoData.

Federated queries

SPARQL queries that retrieve data from remote SPARQ endpoints are called federated queries. The functionality to access the data stored in a remote triplestore and exposed via a SPARQ endpoint on the Web was introduced as an extension to the original SPARQL 1.0 standard and was released as 1.1 version of the standard.

Parliament triplestore, used in this work as a local datastore, does not include a native query processor. In contrast, it utilizes a third-party query processor and SPARQL endpoint, Jena and Joseki respectively. Nevertheless, it is compatible with SPARQL 1.1 standard. Figure 11 illustrates the environment where SPARQL queries are executed.

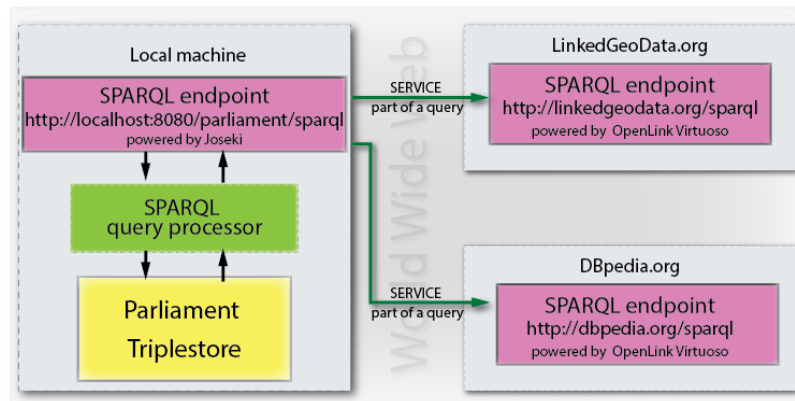


Figure 11. Computational environment of the proof of concept.

In this environment, a local RDF dataset is queried and the retrieved results are joined with data returned from a remote endpoint (DBpedia and LGD in the Figure). To instruct a federated query processor about which portion of a SPARQL query to be invoked against a remote SPARQL endpoint the SERVICE keyword is used. Therefore, a query is composed of two parts. The first part of a query contains a set of triple patterns that are matched to a local graph, when the second part followed the SERVICE keyword includes triple patterns to be matched via a remote endpoint.

Listing 7 presents a snippet that retrieves all the triples relevant to the spatial objects mentioned in the reports. The example of retrieved results are presented in Figure 12.

Listing 7. SPARQL query for selection of the triples related to a LGD object.

```
1 SELECT DISTINCT
2 ?lgdobj ?p ?o
3 WHERE
4 {
5   ?s dcterms:spatial ?lgdobj.
6   SERVICE <http://linkedgeo.org/sparql>
7   {
8     ?lgdobj ?p ?o.
9   }
10 }
```

lgdobj	p	o
http://linkedgeodata.org/triplify/node1258424068	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://geovocab.org/spatial#Feature
http://linkedgeodata.org/triplify/node1258424068	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://linkedgeodata.org/meta/Node
http://linkedgeodata.org/triplify/node1258424068	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://linkedgeodata.org/ontology/Amenity
http://linkedgeodata.org/triplify/node1258424068	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://linkedgeodata.org/ontology/FuelStation
http://linkedgeodata.org/triplify/node1258424068	http://purl.org/dc/terms/modified	"2011-06-05T21:22:48+02:00" ^^< http://www.w3.org/2001/XMLSchema#dateTime >

Figure 12. Example results retrieved by the query in Listing 7.

Inserting data into a triple store

The next step was to construct an INSERT query that would insert selected triples in the triplestore. This operation identifies data with the WHERE clause, which is used to compute solution sequences of bindings for a set of variables. Listing 8 shows this INSERT query. However, it should be mentioned, that with the query in Listing 8, only those values that are directly bound to the spatial objects are retrieved. However, information about object geometries represented as Well-Known text are located one step further. Listing 9 shows the INSERT query that retrieves geometry as WKT and adds this information to the triplestore.

Listing 8. INSERT query

```

1 PREFIX dcterms: <http://purl.org/dc/terms/>
2
3 INSERT
4   {?p ?o}
5 WHERE
6 {
7   ?s dcterms:spatial ?lgdobj.
8 SERVICE <http://linkedgeodata.org/sparql>
9   {
10    ?lgdobj ?p ?o.
11   }
12 }
```

Listing 9. INSERT query selecting geometries


```

1 PREFIX ogc: <http://www.opengis.net/ont/geosparql#>
2 PREFIX dcterms: <http://purl.org/dc/terms/>
3 PREFIX geo: <http://geovocab.org/geometry#>
4
5 INSERT
6   {?geom ogc:asWKT ?o}
7 WHERE
8 {
9   ?s dcterms:spatial ?lgdobj.
10  ?lgdobj geo:geometry ?geom.
11 SERVICE <http://linkedgeodata.org/sparql>
12   {
13    ?geom ogc:asWKT ?o.
14   }
15 }
```

Retrieval of DBpedia data

The next query accessed DBpedia entities via spatial relations. The DBpedia SPARQL endpoint is powered by Openlink Virtuoso. This software supports a number of spatial functions and predicates for the representation of geospatial data. However, these capabilities are not compliant with GeoSPARQL standard; instead, Openlink has developed its own spatial extension to RDF and SPARQL. This creates a source of interoperability problems between spatial functions implemented in Parliament and Openlink Virtuoso. This is especially the case in federated queries, when a portion of a SPARQL query with GeoSPARQL functions is executed against remote SPARQL endpoint run by Virtuoso. In order to mitigate possible misunderstanding between Parliament and Virtuoso endpoints in spatial queries, it has been decided to install Openlink Virtuoso Universal Server to test available spatial access methods.

Section 2.2 explained the spatial content of DBpedia. In general, almost every populated place described on Wikipedia has a geographic reference to a particular point on Earth. Therefore, an article about a city can be found based on the location of the city. For instance, Figure 13 shows a part of description of Lota, a city in Chile, on DBpedia. It is clear that the location of Lota is described as a point with coordinates: -37.0833 (Latitude) and -73.1667 (Longitude).



	<ul style="list-style-type: none">▪ http://da.dbpedia.org/resource/Lota▪ http://sh.dbpedia.org/resource/Lota_(Čile)▪ http://vi.dbpedia.org/resource/Lota▪ http://vo.dbpedia.org/resource/Lota_(Cilän)▪ http://war.dbpedia.org/resource/Lota▪ http://zh.dbpedia.org/resource/洛塔_(智利)
<code>geo:geometry</code>	▪ POINT(-73.1667 -37.0833)
<code>geo:lat</code>	▪ -37.083332 (xsd:float)
<code>geo:long</code>	▪ -73.166664 (xsd:float)

Figure 13. Description of Lota on DBpedia.

Virtuoso supports spatial search based on mutual location of objects. In other words, it is possible to search for entities in some vicinity to a given point. For example, report 4349 features Hospital de Lota, which is the central hospital of Lota. The location of the hospital is known from LinkedGeoData. It seems logical to search for a DBpedia article describing Lota in some proximity to the hospital. Figure 14 provides an illustration for this case. As can be seen from the Figure, the coordinates of Lota point to a place located in less than 2 km from the hospital.

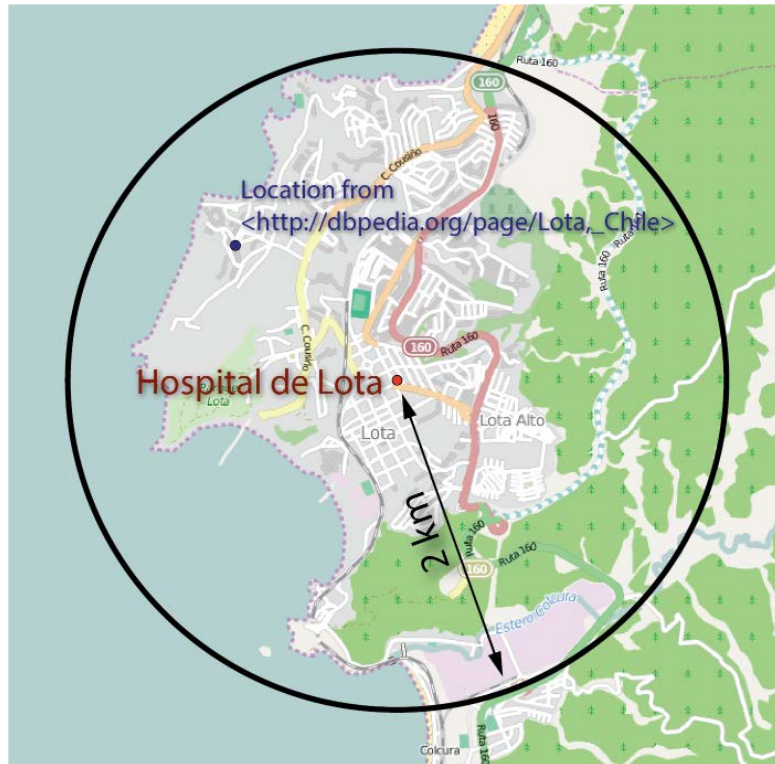


Figure 14. 2 km proximity to Hospital de Lota.

Listing 10 presents a query that selects all the DBpedia entries describing cities in 2 km proximity to spatial objects from Ushahidi reports. Function *bif:st_intersects* has been used in the query. This function has three arguments: geometry of a given point, geometry of candidate objects and a radius around the given point.

Listing 10. Selection of DBpedia entries about cities in 2 km proximity to a report

```

1 PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
2 PREFIX lgdo: <http://linkedgeodata.org/ontology/>
3 PREFIX geom: <http://geovocab.org/geometry#>
4 PREFIX geos: <http://geovocab.org/spatial#>
5 PREFIX scem: <http://schema.org/>
6 PREFIX dbo: <http://dbpedia.org/ontology/>
7 PREFIX dbr: <http://dbpedia.org/resource/>
8 PREFIX wgs: <http://www.w3.org/2003/01/geo/wgs84_pos#>
9
10 SELECT DISTINCT
11 ?s ?label ?ds ?labelDB
12 WHERE
13 {
14   ?s
15     a geos:Feature;
16     rdfs:label ?label;
17     geom:geometry [geo:asWKT ?g].
18
19 SERVICE <http://dbpedia.org/sparql>
20
21 {
22   ?ds
23     a scem:City;
24     rdfs:label ?labelDB;
25     dbo:country dbr:Chile;
26     wgs:geometry ?gd.
27 }
28
29 Filter (bif:st_intersects (?g, ?gd, 2))
30
31 } Order by ?s

```


This query can be used to access information available for a city in DBpedia. Listing 11 provides a variant of this query that retrieves city population, area and mayors' names. This query uses the *OPTIONAL* keyword to indicate optional triple patterns. This is done in order to allow information to be added to the query solution where the information is available, but do not reject the solution because some part of the query pattern does not match. For example, not all the cities presented in DBpedia include information about mayors and without *OPTIONAL* keyword, those cities lacking such information would be rejected completely. Listing 12 gives an example of INSERT query used to add triples from DBpedia to the local triplestore.

Listing 11. Query with *OPTIONAL* keywords

```

1 PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
2 PREFIX lgdo: <http://linkedgeodata.org/ontology/>
3 PREFIX geom: <http://geovocab.org/geometry#>
4 PREFIX geos: <http://geovocab.org/spatial#>
5 PREFIX scem: <http://schema.org/>
6 PREFIX dbo: <http://dbpedia.org/ontology/>
7 PREFIX dbr: <http://dbpedia.org/resource/>
8 PREFIX dbprop: <http://dbpedia.org/property/>
9 PREFIX wgs: <http://www.w3.org/2003/01/geo/wgs84_pos#>
10
11 SELECT DISTINCT
12 ?s ?label ?ds ?labelDB
13 WHERE
14 {
15     ?s
16         a geos:Feature;
17         rdfs:label ?label;
18         geom:geometry [geo:asWKT ?g].
19
20 SERVICE <http://dbpedia.org/sparql>
21
22 {
23     ?ds
24         a scem:City;
25         rdfs:label ?labelDB;
26         dbo:country dbr:Chile;
27         wgs:geometry ?gd.
28 OPTIONAL {?ds dbprop:leaderName > ?leader}
29 OPTIONAL {?ds dbprop:populationTotal ?popt}
30 OPTIONAL {?ds dbo:areaTotal ?area}
31 }
32 Filter (bif:st_intersects (?g, ?gd, 2))
33 } Order by ?s

```


Listing 12. INSERT query for DBpedia triples

```

1 PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
2 PREFIX lgdo: <http://linkedgeodata.org/ontology/>
3 PREFIX geom: <http://geovocab.org/geometry#>
4 PREFIX geos: <http://geovocab.org/spatial#>
5 PREFIX scem: <http://schema.org/>
6 PREFIX dbo: <http://dbpedia.org/ontology/>
7 PREFIX dbr: <http://dbpedia.org/resource/>
8 PREFIX dbprop: <http://dbpedia.org/property/>
9 PREFIX wgs: <http://www.w3.org/2003/01/geo/wgs84_pos#>
10
11 INSERT
12 {
13   ?ds wgs:geometry ?gd;
14     rdfs:label ?labelDB;
15     dbprop:leaderName ?leader;
16     ?ds dbprop:populationTotal ?popt;
17     ?ds dbo:areaTotal ?area.
18 }
19 WHERE
20 {
21   ?s
22     a geos:Feature;
23     rdfs:label ?label;
24     geom:geometry [geo:asWKT ?g].
25
26 SERVICE <http://dbpedia.org/sparql>
27
28 {
29   ?ds
30     a scem:City;
31     rdfs:label ?labelDB;
32     dbo:country dbr:Chile;
33     wgs:geometry ?gd.
34 OPTIONAL {?ds dbprop:leaderName ?leader}
35 OPTIONAL {?ds dbprop:populationTotal ?popt}
36 OPTIONAL {?ds dbo:areaTotal ?area}
37 }
38 Filter (bif:st_intersects (?g, ?gd, 2))
39 }

```

This section has overviewed the construction of the proof of concept. Data initially download from Ushahidi API has been converted into RDF using MOAC vocabulary to encode the report categories. Ontology mapping between Ushahidi categories and MOAC classes helped in this conversion. After that, the data has been manually enriched with links to relevant entities in LinkedGeoData. *Section 4.2* has shown how to access additional information from DBpedia and LinkedGeoData using federated SPARQL queries.

Results as well as their evaluation are given in the following section.

Chapter 5. Evaluation of the results

This chapter presents the findings and reviews the experience obtained during the construction of the proof of concept. The chapter is structured following the same division of the work as in Chapter 4. Thus, the section starts with the evaluation of the first work package and then overviews the outcomes of the semantic enrichment achieved in the second package. Comparison of data management techniques applicable to the data before and after the semantic enrichment is used to evaluate to what extent such an enrichment can help decision makers to answer queries. In addition, several use cases are used to illustrate practical value of the achieved semantic enrichment for decision makers involved in EM.

5.1 Evaluation of work package 1

Briefly, the first work package dealt with the conversion of Ushahidi data from the tabular format into RDF. The conversion was done using the MOAC vocabulary to represent categories of the reports.

Ontology mapping between the Ushahidi categories and the MOAC terms proved considerable degree of suitability of the vocabulary for representation of Ushahidi categories. Appendix C provides a table showing a correspondence between two systems of categories used in Chile 2010 and Haiti 2010 and the MOAC vocabulary. This table clearly indicates that the MOAC core vocabulary makes it possible to express semantics of the Ushahidi reports. Despite of the fact, the vocabulary lacked 14 out of total 48 categories used for the Chile deployment (see Appendix B), absent classes was easily substituted by terms and notions from other vocabularies. Another solution to the lack of required notions was to extend the vocabulary with absent classes.

Another candidate vocabulary - The Humanitarian Exchange Language (HXL) was put to the test. However, due to the fact it did not possess terms related to categories of Ushahidi, the use of the MOAC was concluded to be more desirable than the HLX vocabulary.

Another advantage of the MOAC was in the structure of the vocabulary (see Section 3.3). It was shallow and very easy for understanding. This made MOAC to be well understood by people from outside the EM domain. Therefore, semantic interoperability of Ushahidi data could be increased by providing RDF-based ontology mapping between Ushahidi categories and MOAC types in the onset of a disaster. In such a way, countless stakeholders involved in a disaster relief operation could share common understanding of incidents captured in MOAC.

If MOAC provided suitable terms for description of categories, the Dublin Core vocabulary and The SIOC Core Ontology were used to describe general attributes of a report – an incident title, incident date, location and content. These terms are widely used in many domains since they reflect very common properties of information. In addition, the terms are quite self-explanatory, which makes them easy to use by general public.

As it was explained in *Section 4.2* Ushahidi data had considerable amount of misspellings and typos in location description, not to mention ambiguity of georeferencing. Eighty one reports were enriched with links to 95 spatial features mentioned in the reports. As a result, the reports received explicit descriptions of incident locations. Therefore, the use of LGD URIs for locations not only provides additional means for georeferencing, but also mitigates one of the problems of VGI - different people name the same things differently. In other words, LGD URIs provide shared and formalized way to describe an object on Earth.

To sum up, the first work package revealed general suitability of MOAC vocabulary to support the representation of incident categories in RDF-based Ushahidi data. Another outcome of this part of the

work was the understanding of the role of LGD URIs as a universal and explicit way to point at a particular spatial feature.

5.2 Evaluation of work package 2

The second work package was aimed at the development of SPARQL queries to retrieve relevant information from LinkedGeoData and DBpedia. Two different approaches to access data were used. In the first approach, information linked to the identified objects was retrieved from the LGD SPARQL endpoint. The second approach was performed by accessing DBpedia data using spatial relations between entities in DBpedia and LGD. In both cases, federated queries were used.

Enrichment with LGD data

The general idea behind the queries against LGD data was to retrieve all the triples having the identified URIs as subjects. However, the query in Listing 7 retrieved information related to 64 URIs out of total 95. The remaining 31 URIs, mostly roads and bridges, were unknown by the SPARQL endpoint of LGD. The reason for this was the fact that LGD, as a project, was aimed at providing easy mechanism for accessing RDF-based representation of OSM data. However, for the sake of performance, only a limited number of objects were exposed via the SPARQL endpoint. Nevertheless, the entire database of OSM could be accessed as Linked Data via the LGD API. An RDF description of lacking 31 objects were downloaded from the API and were inserted into the triplestore using local SPARQL endpoint. The list of missing objects can be seen in Appendix D.

Another issue occurred while dealing with the spatial data retrieved via a Virtuoso powered endpoint. Such an endpoint returned geometries using a not standardized representation of WKT data implemented in Virtuoso. The query in Listing 13 with retrieved results in Figure 15 make it clear that those 31 objects accessed via the LGD API had a different datatype than the 64 objects obtained via the LGD endpoint. For the sake of interoperability, it is worth to change the datatype to the standardized one - <http://www.opengis.net/ont/GeoSPARQL#WKTLiteral>. The problem was solved by rewriting the datatype of the geometries. The query for this is shown in Listing 14.

Listing 13. Query to check a datatype of geometry representation.

```
1 SELECT DISTINCT
2 (Count(?s)) ?datatype
3 WHERE {
4   ?s a geos:Feature;
5     geom:geometry [
6       geo:asWKT ?lgdgeom].
7 BIND(datatype(?lgdgeom) as ?datatype)
8 }
9 Group by ?datatype
```

.1	datatype
"64"	http://www.openlinksw.com/schemas/virtrdf#Geometry
"31"	http://www.opengis.net/ont/geosparql#wktLiteral

Figure 15. Results retrieved by the query in Listing 13.

Listing 14. Query to change the datatype of a geometry representation.

```
1 DELETE { ?sgeom geo:asWKT ?lgdgeom }
2 INSERT { ?sgeom geo:asWKT ?lgdgeomwkt }
3 WHERE
4 {
5   ?s a geos:Feature;
6     geom:geometry ?sgeom.
7   ?sgeom geo:asWKT ?lgdgeom.
8   BIND(datatype(?lgdgeom) as ?datatype)
9   Filter (datatype(?lgdgeom) = <http://www.openlinksw.com/schemas/virttrdf#Geometry>)
10  BIND(STRDT(STR(?lgdgeom), geo:wktLiteral) AS ?lgdgeomwkt)
11 }
```

Limited availability of data via an endpoint creates a significant source of inconvenience. For instance, when a person issues a federated query against such an endpoint, he or she receives an incomplete set of solutions. The problem is that it is difficult to tell whether the incompleteness are caused by absence of queried data in the database or there are restrictions on the amount of data that can be retrieved via federated queries. The latter can be seen as a, so-called, black box problem. Particular adjustments of a remote SPARQL endpoint are often unknown; and the amount of retrieved data can be restricted in order to prevent a server overload. The most obvious solution to this is to have a local version of a database of interest.

Enrichment with DBpedia data

Spatial access methods were used to retrieve DBpedia data related to the settlements where identified objects were located.

The query in Listing 10 retrieved solutions only for 30 objects out of the total 95. These 30 objects were points when the rest 65 were linestrings. The problem with the retrieval of linestrings was caused by the fact that version of Virtuoso server (version 7.1) used in the construction of the proof of concept supported spatial relations only between point objects. This drawback eliminated possibility to retrieve any information related to objects with more complex geometries than points. This problem could be solved in 2 different ways. The first way implied a calculation of the centroids for the linestring objects to use them instead of linestring objects. However, this approach required the calculation of centroids outside the database with further inserting of them into the database with proper predicates. This was due to the fact that GeoSPARQL standard lacked functionality for calculation of centroids by a database. Another approach was to retrieve description of all the Chilean cities presented in DBpedia, as has been done. Having this information in a triple store supporting GeoSPARQL makes it possible to use a wider and richer functionality of available spatial access methods than in Virtuoso.

Summing up, informational retrieval from remote SPARQ endpoints is a quite tricky task. On the one hand possible restrictions of the size of retrieving data humpers the process of query debugging. Unexpected solutions can be produced by a correct query as a result of such a restriction. On the other hand, the fact that not the entire data set is exposed via an endpoint also complicates the exploration of available data. These two restrictions create an urgent need for having local copy of candidate datasets. Such an approach allows avoiding possible interoperability issues that arise between implementations of spatial functions in different datastores. Moreover, many LOD resources use Virtuoso powered endpoints with limited and unstable support of spatial access methods.

5.3 Evaluation of emerged data management techniques

Multi criteria filtering of incidents

The Ushahidi platform provides a limited number of data management techniques. The user interface of Ushahidi allows only filtering of reports based on their categories. However, the selection of reports can only be performed using a single category. A selection based on multiple categories is not possible. Figure 17 shows a screen dump with the interface of the Chile Ushahidi deployment. As can be seen from the Figure, there is a list of categories each of which can be selected separately. Because of this selection, reports assigned with the chosen category appears on the map.

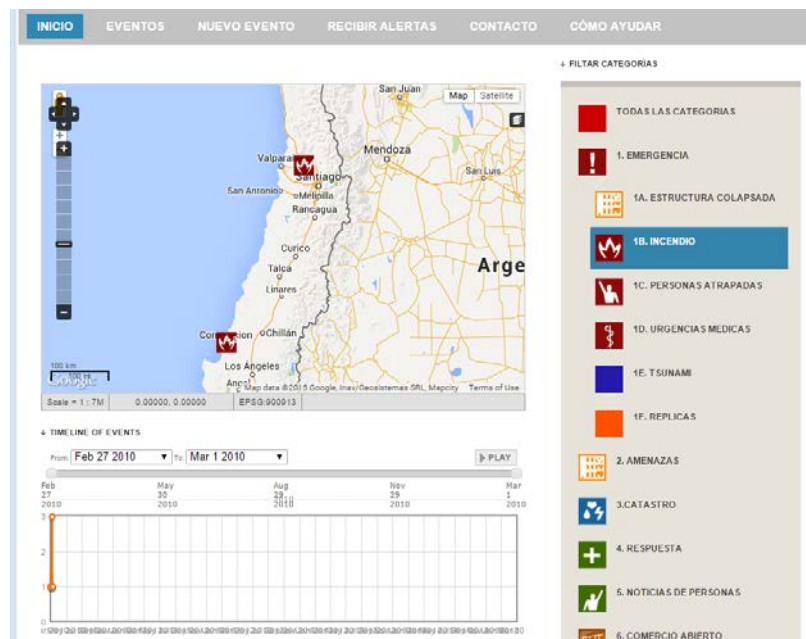


Figure 16. Interface of Chile Ushahidi deployment.

In contrast, triplification of the Ushahidi reports using the MOAC vocabulary to represent the categories makes it possible to filter the reports using multiple categories. This functionality allows prioritizing of the reports based on a set of categories. Listing 4 gives an example of a query that counts the number of categories assigned to a report. Based on the results of such a query it is easy to select those reports that describe incidents with more complex semantics.

Interoperable georeferencing and spatial querying

Semantic enrichment obtained from LGD allowed explicit georeferencing of the reports in an interoperable way. One of possible use cases is to extract object geometries for further use in information systems. For instance, geometries of blocked roads can be accessed and transferred to navigation systems used by humanitarian agencies to solve routing problems. Moreover, geometries are represented using well-known text (WKT) format, which is an interoperable and a widely used format defined by the ISO/IEC 13249-3:2011 standard. Listing 15 shows how to extract geometries of blocked roads, when Figure 17 provides example results retrieved by the query. In the initial state meaningful manipulation with spatial information was almost impossible.

Listing 15. Selection of blocked roads and their geometries.

```

1 SELECT DISTINCT ?s ?lgdobj ?g ?name
2 WHERE
3 {
4   ?s dc11:subject MOAC:RoadBlocked;
5     dcterms:spatial ?lgdobj.
6   ?lgdobj geom:geometry [
7     geo:asWKT ?g].
8   OPTIONAL {?lgdobj rdfs:label ?name }
9 }

```

s	lgdobj	g	name
http://localhost:3333/chile/4282	http://linkedgeodata.org/triplify/way2726459	<code>"LINESTRING(-72.6922017 -36.4666835 -72.696655 -36.467567200000005 -72.7005721 -36.4682427)"</code>	
http://localhost:3333/chile/4280	http://linkedgeodata.org/triplify/way144276507	<code>"LINESTRING(-73.4418546 -37.2522754 -73.4438537 -37.2514486)"</code>	"Arauco - Llico"
		<code>"LINESTRING(-70.69446400000001 -33.3429191 -70.69449680000001 -33.343147 -70.6944994 -33.3432746 -70.6944952 -33.343564 -70.6944266 -33.3442616 -70.69437640000001 -33.3450304 -70.6942959 -33.3461501 -70.69418660000001 -33.3473206 -70.6940811 -33.3484258 -70.6939895 -33.349406300000005 -70.6939341 -33.3498092 -70.6938499 -33.350252600000005 -70.6937667 -33.350650300000005 -70.693664 -33.350979200000005 -70.6935344 -33.3513685 -70.69337 -33.351757400000004 -70.6931339 -33.352254800000004 -70.6929513 -33.352665200000004)"</code>	"Autopista Los Libertadores"
http://localhost:3333/chile/4268	http://linkedgeodata.org/triplify/way25422270	<code>"^<http://www.opengis.net/ont/geosparql#wktLiteral>"</code>	

Figure 17. Example results retrieved by the query in Listing 15.

Object geometries are not the only information available in LGD. Attributive values and a classification of objects provided by the LGD ontology are also sources of additional information. For instance, bridge descriptions often include information about a number of lines, a type of the pavement, capacity, etc. However, the presence of additional attributive values are not consistent. Richness of available information significantly varies depending on the location of an object. In general, objects situated in areas with high population density have a better description in comparison with objects in remote areas. In addition, significant and outstanding objects, for instance the main bridge of a city, also described with more details than ordinary objects. This peculiarity of the content distribution was inherited from the crowdsourced nature of the OSM data. The more people live in a neighborhood the greater the chance it is mapped.

As a result of such uneven geographical distribution of the content, the use of *optional* clauses (see Section 4.2) are preferred in order to avoid rejection of the solutions with missing *optional* triple patterns. Listing 16 shows a query selecting bridges located in 10-kilometer proximity to compromised bridges. It shows object labels and type of pavement, however, they are *optional*. This query uses *geof:sfContains()* GeoSPARQL functions. This is a Boolean function, which returns “true” if a geometry of the first argument contains a geometry of the second argument. In the query, *geof:sfContains()* checks whether a 10-km buffer (variable “?badbridgebuffer”) constructed around a compromised bridge (variable “?badbridgegeom”) contains a geometry of an undamaged bridge (variable “?goodbridgegeom”). A buffer is built using the *geof:buffer()* function, which has 3 argument: a geometry, a size of a buffer, units expressing the buffer size.

One can notice that such a query is an example of a k-nearest neighbor (kNN) query. kNN queries select k nearest objects to a given location. However, GeoSPARQL standard does not natively support spatial kNN queries (Patroumpas et al., 2014b). Because of this, a spatial range query is used in Listing 16. In other words, it selects all the objects located in a given area (?badbridgebuffer in the query) and then orders the results based on the distance between the geometry of a given object (?badbridgegeom) and candidate geometries (?goodbridgegeom).

Listing 16. Selection of bridges located in 10-km proximity to compromised bridges.

```

1 SELECT DISTINCT
2     ?badbridge
3     ?badbridgelabel
4     ?goodbridge
5     ?goodbridgelabel
6     ?surface
7     (geof:distance(?badbridgegeom,?goodbridgegeom, units:metre) as ?distance)
8 WHERE
9 {
10     ?report a MOAC:UshahidiReport;
11     dc11:subject MOAC:CompromisedBridge;
12     dc11:title ?reporttitle;
13     dcterms:spatial ?badbridge.
14
15     ?badbridge geom:geometry [
16         geo:asWKT ?badbridgegeom
17     ].
18 OPTIONAL {
19     ?badbridge rdfs:label ?badbridgelabel
20 }
21
22     ?goodbridge lgdo:bridge "true"^^xsd:boolean;
23
24     geom:geometry [
25         geo:asWKT ?goodbridgegeom
26     ].
27 OPTIONAL {
28     ?goodbridge rdfs:label ?goodbridgelabel.
29     ?goodbridge lgdo:surface ?surface.
30 }
31 BIND(geof:buffer(?badbridgegeom, "10000"^^xsd:double, units:metre) as ?badbridgebuffer)
32 Filter (geof:sfContains(?badbridgebuffer,?goodbridgegeom))
33 } Order by ?badbridge

```

The approach demonstrated in Listing 16 can be used to select any objects that are spatially related. Examples include any queries that select objects of a particular type located nearby with any available *optional* information.

Listing 17. Selection of officials of populated places where operating hospitals are located.

```

1 SELECT DISTINCT
2 ?report ?reporttitle ?lgdlabel ?db ?leadername
3 WHERE
4 {
5 ?report a MOAC:UshahidiReport;
6     dc11:subject MOAC:HospitalOperating;
7     dc11:title ?reporttitle;
8     dcterms:spatial ?lgdobj.
9
10 ?lgdobj geom:geometry [
11     geo:asWKT ?lgdgeom].
12 OPTIONAL {?lgdobj rdfs:label ?lgdlabel}
13
14 ?db dbo:type dbr:Communes_of_Chile;
15     w3geo:geometry ?citygeom.
16 OPTIONAL {?db dbprop:leaderName ?leadername}
17
18 BIND(geof:buffer(?lgdgeom, "2000"^^xsd:double, units:metre) as ?lgdbuffer)
19 Filter (geof:sfContains(?lgdbuffer,?citygeom))
20
21 } Order by ?report

```


Background knowledge for better response

Semantic enrichment of Ushahidi data with additional information from DBpedia brings valuable informational background for the initial content of the reports. For instance, quick access to names of officials can be of importance for people involved in decision-making. Listing 17 provides such a query. It selects names of officials of populated places where operating hospitals are located. The same approach as in Listing 16 is used. Objects are selected based on a mutual location and an additional thematic information is bound to the solutions.

However, folksonomic nature of DBpedia ontology should be always kept in mind. Semantically similar notions are often presented in different lexical forms across the ontology. For instance, a head of a city is termed as a mayor in English speaking countries and an “alcalde” in countries with mostly Spanish speaking population. The possible solution to this problem is quite similar to the one related to uneven geographical distribution – to use *optional* triple patterns featuring all identified variants.

DBpedia provides broad descriptions of resources. Literally, every piece of information presented on Wikipedia can be found in DBpedia. This allows calculation of derivative characteristics based on available data. For instance, general information about the size of an area or a population of a city can be used for calculation of population density. In turn, population density can be used for prioritizing of response actions. Incidents occurred in areas with a higher population density may require a quicker response than incidents in rural areas. Listing 18 gives a query that prioritizes reports based on the density of population.

Listing 18. Prioritizing of the reports based on the density of population.

```
1 SELECT DISTINCT
2     ?report
3     ?reporttitle
4     ?cityname
5     ?density
6     (geo:distance(?badbridgegeom,?goodbridgegeom, units:metre) as ?distance)
7 WHERE
8 {
9     ?report a MOAC:UshahidiReport;
10         dc11:subject MOAC:RoadBlocked;
11         dc11:title ?reporttitle;
12         dcterms:spatial ?lgdobj.
13
14     ?lgdobj geom:geometry [
15         geo:asWKT ?lgdgeom].
16
17     ?db dbo:type dbr:Communes_of_Chile;
18         w3geo:geometry ?citygeom;
19         rdfs:lable ?cityname.
20     OPTIONAL {?ds dbprop:populationTotal ?popt}
21     OPTIONAL {?ds dbprop:areaTotal ?area}
22
23     BIND((?popt/?area) as ?density)
24     BIND(geof:buffer(?lgdobj, "2000"^^xsd:double, units:metre) as ?lgdbuffer)
25     Filter (geof:sfContains(?lgdbuffer,?citygeom))
26
27 } Order by ?report ?density ?distance
```

This sub section has shown that representation of Ushahidi reports as Linked Data significantly widens the range of data management techniques applicable to the data. In the initial state only

information about categories and geographic coordinates can be used. The other attributes are very inconsistent and ambiguous. Representation of categories as MOAC terms allows multiple criteria filtering. The use of LGD URIs enhances georeferencing of the reports as well as provides more robust framework for collaborative work of volunteers. In addition, it gives quick access to the geometries of objects in an interoperable way. Moreover, additional information can be accessed via spatial queries. In such a case, *optional* thematic information can be bound to solutions retrieved via spatial range queries. Background information from DBpedia helps to understand the context of incidents.

Chapter 6. Discussion, Conclusions and Recommendations.

This chapter brings together all the information acquired from the literature review (Chapters 2 and 3) and practical outcomes obtained during the construction of the proof of concept (Chapters 4 and 5). The chapter starts with a discussion on the findings of Chapter 5. The discussion critically overviews methods and approaches used in the work giving an understanding of limitations emerged as a result of the implementation. Conclusions are stated in the following subsection after the discussion. Conclusions are structured according to the same logic of research questions in *Section 1.4*. After the conclusions, the last chapter subsection introduces a set of recommendations for future research in the topic of this thesis.

6.1 Discussion

Integration with LOD mitigates human factor

Sections 4.1 - 4.2 and 5.1 proved the assumption that an implementation of a single and agreed upon vocabulary allowed increased semantic interoperability of the Ushahidi reports. The MOAC classes and properties were self explanatory and the vocabulary itself was easy to understand. The latter is an important characteristic, since Ushahidi volunteers might lack deep knowledge of the EM domain terminology. Therefore, dissemination of ontology mapping between Ushahidi categories and the MOAC terms could help in communication of knowledge between all the parties involved. In addition, MOAC could easily be extended with new classes if they were needed for a particular deployment.

The use of the LinkedGeoData URIs as universal links to locations eliminate the possibility to point at a location in an ambiguous way. There was less room for a mistake in providing one single URI of a place than in writing down its full name or address. In this sense, the LGD project proved its importance for Linked Data initiatives as a source of spatial dimension for the Web of Data. Moreover, additional semantics obtained from structured information presented in LGD allowed quick access to very useful background knowledge about infrastructural objects and public buildings, which could be utilized by decision makers.

SPARQL endpoints and GeoSPARQL

Sections 4.3 and 5.2 described the development of SPARQL queries to retrieve relevant to the reports information from the SPARQL endpoints of LinkedGeoData and DBpedia. Encountered issues (constraints on the number of triples to be retrieved and an exposure of the limited dataset) made an idea of dealing with federated queries against those endpoints to be questionable. This led to an understanding, that, for the sake of convenience, it was highly desired to have a copy of interested data in a local triplestore. Moreover, both DBpedia and LinkedGeoData provided an option to download their entire databases. However, it worth mentioning that the size of the data dumps was considerably voluminous, which could potentially create additional problems with bulk loading and the performance of a triple store. This thesis did not go further with exploration of possible problems related to work with data dumps of DBpedia and LGD; the proof of concept dealt with a very limited subset of Ushahidi reports. Therefore, it was more preferable to obtain interested data manually via APIs and SPARQL endpoints rather than to work with giant data dumps.

Geospatial capabilities of triplestore implementations used in the work had not reached their maturity not to mention interoperability problems that arose between software of different brands. Support of geospatial data and access methods in Virtuoso is still under development, different versions of software interpreted the scope of spatial functions and predicates differently. Moreover,

Virtuoso did not support GeoSPARQL functions implemented in Parliament; as a result, the use of spatial federated queries issued between those systems was not possible.

Even though, Parliament had full support of GeoSPARQL, the standard itself lacked one of the basic functions, namely, selection of k-nearest neighbors. In the work, kNN was simulated using combination of a spatial range query with following ordering of results based on a distance to given objects (see Listing 16). Clearly, such an approach created additional computational workload, which could have been avoided if a kNN algorithm was used. For instance, several proprietary software products such as Oracle Spatial and Graphs have an explicit kNN search.

Improved capability to answer “Where” questions

RDF as a framework provided flexible mechanism for data modeling. Conversion of the reports into RDF created an opening to change the cardinality of relationships between reports and their attributes to one-to-many (see Figure 5.). This resulted in a finer granularity of the data, which in turn, led to more precise georeferencing of incidents. Thus, the change of cardinality released considerable amount of location information, which was trapped in the initial data model.

The enhanced data model, together with WKT geometries obtained from LGD, considerably widened the range of possible spatial queries to be run against the data (see *Section 5.3*). Geometries, as explicit location descriptions allowed accessing additional semantics via spatial relations, for instance from DBpedia. This provided semantic enrichment with facts about officials, population, area, etc. Moreover, WKT representation allowed better interoperability of spatial information between systems.

6.2 Conclusions

This section presents conclusions drawn from the entire work. The first subsection gives an answer to the main research question. After that, the answers to the sub research questions are presented.

6.2.1 Main conclusion

This thesis was set out to investigate ***to what extent the Linked Open Data cloud could help to semantically enrich volunteered geographic information in order to better answer queries in the context of crisis and disaster relief operations.*** In general, this question implied an identification of obstacles related to querying of VGI in the context of crisis and disaster relief operations and then figuring out how such obstacles could be overcome with a semantic enrichment obtained from the LOD cloud.

Section 3.1 showed a number of stakeholders involved in a disaster, which reflected in a significant variety of possible use case scenarios applicable to the data. Therefore, it was concluded that the requirements for data use in the context of crisis and disaster relief operations were rather general rather than domain-specific.

In contrast to this, the content of the Ushahidi data was domain-specific and required the use of dedicated vocabularies to express the semantics of the reports. The MOAC vocabulary helped to encode into RDF the semantics of pleas for help (*Sections 4.2 and 5.1*). As a result, it became possible to apply multi criteria filtering to the reports based on their semantics (*Section 5.3*).

The Ushahidi data was plagued with drawbacks common to most user-generated content (analyzed in *Section 4.1-4.2 and 5.1*). Inherent in VGI messiness and inconsistency of the data, together

with ambiguous georeferencing hampered consistent data management. Because of this, valuable EM information was locked in the initial data set.

In RDF, there is not any limitation on the number of objects connected to a subject. This allowed storing more than one location per a report, thus, solving the georeferencing problem. The semantic enrichment achieved with the LGD data increased interoperability of spatial content and gave a robust spatial dimension to the reports. These improvements made it possible to access DBpedia entities via spatial relations. As a result, comprehensive queries could be constructed. For instance, it became possible to prioritize the reports based on the density of population or to extract useful information about local amenities, official names, infrastructural objects, etc.

This thesis practically proved that integration of VGI with relevant entities in the LOD cloud made it possible to semantically enrich unstructured user-generated content with structured information presented in LOD. The LOD cloud can be perceived as an informational skeleton. Scattered blobs of unstructured data, being attached to this skeleton, acquire an integrated dataspace where a standardized navigation can be used. Despite of the fact, the work dealt with the disaster-related VGI, the demonstrated approach can be transferred to other VGI if there is a need for better handling of the data.

6.2.2 Answering sub research questions

1. *What standards and tools facilitate semantic integration of disaster related crowdsourced information?*

Chapter 2 thoroughly overviewed the technologies facilitating the semantic integration of disaster related crowdsourced information. As described in *Section 2.1*, there are several specific standards, that made the existence of the Semantic Web possible. Those standards were Resource Description Framework (RDF) and SPARQL Protocol and RDF Query Language (SPARQL). In addition, the Web Ontology Language (OWL), a vocabulary extension to RDF, was another standard that helped to publish and share ontologies on the World Wide Web. GeoSPARQL was an extension to SPARQL allowing spatial capability into SPARQL queries.

Tools were described in *Section 2.4*. Overall, a significant number of commercial and open source software products could facilitate functionality needed on different stages of data conversion, storage and retrieval. Roughly, tools could be divided into several groups based on the functionality they provide. *Conversion tools (Section 2.4.1)* were meant to help in conversion of datasets from different formats to RDF representation with various serialization. In this thesis, *OpenRefine* was used to convert the data from tabular form into RDF (*Section 4.2*). *RDF generators (Section 2.4.1)* provided access to relational databases as virtual, read-only RDF graphs. *Validators (Section 2.4.2)* checked RDF datasets to prevent users from malformed input. The work utilized the functionality of **RDF Validator** by W3C to validate the dataset obtained after the conversion (*Section 4.2*). *Semantic Web browsers (Section 2.4.3)* helped to browse and navigate through data published as Linked Data on the Web. *Triplestores* were databases built for the storage and retrieval of triples using semantic queries. Two triplestore implementations, Parliament and Virtuoso, were used in the work. *Query builders* provided visual interfaces guiding and assisting in the process of query construction. During the work, iSPARQL was used many time since it was implemented as an interface for LGD and DBpedia endpoints.

2. *What ontologies can be used for data conversion into RDF and how?*

There are many ontologies to support data conversion into RDF. The Linked Open Vocabulary (LOV) project provides structured access to almost all prominent vocabularies on the Web (see Figure 1). In the current work, several ontologies were tested including domain-specific: the Management of a Crisis vocabulary, The Humanitarian eXchange Language, and more general: DublinCore, SIOC and several geo ontologies. Table 2 summarizes the ontologies used in the work and gives an understanding of how they were used. Detailed description of ontologies can be found in *Section 4.1 and 4.2*

3. *What Linked Data Hubs can be used for establishing outgoing links from RDF-based crowdsourced data?*

Section 2.2 gives an overview of the Linked Open Data resources and their geographical content. In the work, semantic links were established to entities of the LGD data (described in *Section 4.2*). Those links provided a connection between descriptions of places featured in the reports and the representation of those places in the LGD database. Links to LGD were the only explicit connections between the data and LOD resources established in the work. DBpedia data was accessed via implicit spatial relationships that existed between LGD and the DBpedia entities.

4. *What is the difference between integration of VGI into the Linked Open Data cloud in the case of poor and rich information environment?*

This question was formulated with the intention to investigate potential obstacles provoked by absence of data to which VGI can be linked. In this work, reports were integrated with LGD data. It became apparent that the LGD data (derived from OSM) could provide all the objects identified from the reports. Moreover, the Ushahidi data featured quite common types of objects such as hospitals, roads, bridges, schools, gas stations etc. These objects were of importance for the society and the OSM project mapped them in the first place. Therefore, it was concluded that in the case of integration of the Ushahidi data with LGD it was possible to find at least a point representation of an object of interest.

5. *What questions are difficult to answer using existing information management techniques in the context of crisis and disaster management?*

Description presented in *Section 5.3* provided an overview of the data management techniques applicable to the data in the initial state and compared them with possibilities emerged after the conversion and the semantic enrichment. To sum up, in the initial state, the range of applicable data management techniques was very limited. Querying of the data was possible only through matching of substrings (a keyword-based search). This approach produced unreliable results due to the weak quality of the data. The web interface of Ushahidi (see Figure 16) allowed only a single-category filtering of the reports. Moreover, ambiguous georeferencing led to the partial loss of location information if a report described several incidents. Therefore, questions like “What reports have categories of “ShelterOffered” and “Distribution of water” and where they are located?” were difficult to answer.

6. *Which of them can be answered by posing GeoSPARQL queries across encoded into RDF disaster related VGI linked to multiple Linked Data resources?*

After the conversion and enrichment the range of data management techniques applicable to the data significantly widen. Implementation of MOAC for category encoding made it possible to count the categories assigned to a report as well as to filter the reports using multi category criteria. This improvement brought possibility to distinguish reports based on the complexity of their semantics. The greater the number of categories assigned to a report the more complex semantic of incidents it

described. Utilization of LGD URIs for location description as well as the emerged possibility to describe more than one location per a report fixed the problem of ambiguous georeferencing, thus, allowing to pose complex spatial queries across the data in the LOD cloud. *Section 5.3* provided the examples of several useful queries to illustrate the potential of answering questions. For instance, Listing 16 demonstrated an approach that could be used to select required objects spatially related to a report with a given semantic. Therefore, a question “What are the locations of operating hospitals, situated in less than 10 km proximity to collapsed schools and what are the names of officials in those places” could be answered.

7. *How to construct a SPARQL query? What tools are able to assist in construction of a SPARQL query for a SPARQL endpoint?*

Section 2.1 briefly overviews the SPARQL technology. A SPARQL query comprises of a prefix declaration, data definition, a result clause, the query pattern and query modifiers. Triple Patterns are written as subject, predicate and object. A query processor matches query graph patterns with those in the data and produces a solution sequence, where each solution has a set of bindings of variables to RDF terms. SPARQL FILTERs are used to restrict solutions to those for which the filter expression evaluates to TRUE. It is useful to be able to have queries that allow information to be added to the solution where the information is available, but do not reject the solution because some parts of the query pattern do not match. Optional matching provides this facility: if the *optional* part does not match, it creates no bindings but does not eliminate the solution.

Several tools existed to assist in construction of a SPARQL query, which are presented in *Section 2.4.6*. In general, most useful for a query construction functionally was to validate queries before running them. Many SPARQL endpoint implementations had such a service as a default capability. It was very useful to know which part of a query had errors. However, despite of the fact that some endpoint implementations also provided a GUI (iSPARQL for instance) allowing the use of predefined types, query forms and predicates from common ontologies, such assistance was concluded to be useless or unnecessary. This conclusion was drawn from the experience of mastering SPARQL queries. The learning curve of SPARQL was quite steep and by the point when the author had understood the structure and mechanism of SPARQL it was not an issue to manipulate with needed prefixes and triple patterns without any assistance.

8. *What is special about spatial SPARQL queries? “How to add a spatial condition into a SPARQL query?” and “What tools can provide required for this functionality?”*

Peculiarity of spatial SPARQL queries was discussed in *Sections 4.3, 5.2, 5.3 and 6.1*. By now, several triplestore implementations allowed spatial capabilities. However, very few of them supported GeoSPARQL standard. In contrast, many software vendors implemented their own vision of functionally, which was incompatible with GeoSPARQL. This conclusion was supported by the experience of using Parliament and Virtuoso triple stores. Parliament had a full support of GeoSPARQL when Virtuoso used its own set of spatial function. This led to interoperability problems with spatial federated queries issued between systems (see *Section 6.1*). In addition, some shortcomings of GeoSPARQL were observed. For instance, this standard did not included k-nearest neighbors selection function which was a significant drawback. However, it is fair to say that Virtuoso also lacked kNN. It also worth mentioning, that Virtuoso received the support of complex geometries only in February of 2015. Therefore, a user should be aware that versions of software before to 7.2 do not support line strings and polygons in many spatial functions.

There are two main ways how to add a spatial condition into a query. The first way is to state spatial conditions in the SELECT clause. This allows counting and ordering of the solutions based on

this spatial condition. Another way is to add restriction of the solutions based on spatial relations via SPARQL FILTERs. This way is illustrated with Listings 9, 10, 11.

Conclusion about the use of tools that help to introduce spatial conditions into SPARQL queries corresponds with the conclusion about usability of more general tools for construction of SPARQL queries given in previous answer. Spatial reasoning in SPARQL queries was straightforward and clear. Therefore, no any specific tools were needed.

9. What tools can provide visualization of the results retrieved by GeoSPARQL queries?

The SPARQL protocol uses an XML schema, which is uncommon for non-semantic web environment. Thus, for the sake of visualization, the returned data was to be transformed into JSON structures and then be visualized with the help of tools. Required functionality can be found in Sgvizler and D3 JavaScript library (see *Section 2.4*).

10. How correct are the results retrieved by GeoSPARQL queries?

Correctness of the results retrieved by a query depends on the trustworthiness of the queried data and the correctness on the query itself.

The first issue, quality and reliability of the data exposed via a remote endpoint are up to the owner of the dataset. In this work, LGD and DBpedia were used to achieve semantic enrichment. Both projects provide a Linked Data version of the data collected by their respective parent organizations, Wikipedia and OpenStreetMap. Therefore, the quality of the data was the same as in the donor datasets. In general, both datasets had shortcomings, such as an inconsistency of information, shallow and unstructured ontologies.

Related concern touches upon the quality of the Ushahidi data. The correctness of the assigned categories was not evaluated and their relevance was taken for granted. However, it should be mentioned that because volunteers assigned the categories errors have emerged due to the human factor.

In the case if a GeoSPARQL query retrieves unexpected results, the spatial component of the query should be checked first. Spatial SPARQL queries consist of two parts, a set of triple patterns describing thematic component and spatial filters. Spatial functions used in filters operate on the object geometries. In the work, the main obstacle was related to the use of the proprietary datatypes for geometries in the remote endpoints. Virtuoso uses `<http://www.openlinksw.com/schemas/virtrdf#Geometry>` for WKT geometries. GeoSPARQL standard, in contrast defines the WKT using `<http://www.opengis.net/ont/GeoSPARQL#WKTLiteral>`. Therefore, a user should be aware about which datatypes are used on a remote endpoint, and be able to transform them in the query.

11. To what extent does difference in richness of informational environment influence the robustness of the retrieved results?

This sub question addresses “*What if there is not required graph pattern due to scanty data? Or what if there is only part of the required pattern?*” In version 1.0 of the SPARQL standard there were no means to query graph patterns with missing parts. Version 1.1 extended the standard introducing the *OPTIONAL* keyword for triple patterns aimed at potentially missing data. When such data is missing the *optional* conditions will not lead to the rejection of the entire query. In this way, a user is always assured that he or she retrieves all the data available in the datastore with indication of missing parts.

6.2 Recommendations

To Open Geospatial Consortium

Undoubtedly, the emergence of the GeoSPARQL standard was a notable milestone in the development of technologies facilitating the Semantic Web. Standardization of spatial vocabularies brought interoperable semantics to spatial queries. Spatial functions implied by this standard provided a useful mechanism to access data via a spatial dimension. However, available functionality did not allow calculations of centroids by a database and selection of k-nearest neighbors. The former is very helpful when it comes to querying remote endpoints supporting only point data. The ability to issue kNN queries mitigates unwanted computational workload that emerges as a result of substituting a kNN with a combination of a spatial range queries followed by ranging of the results. Consequently, it is recommended to extend the standard with these two functions.

To the Ushahidi project

This thesis concludes the Ushahidi data can be significantly improved by the semantic enrichment achieved from LGD and DBpedia. Such an enrichment brings better semantic interoperability of the content as well as enhances spatial and thematic component of the data. Thus, the recommendations are as follows:

- To use the MOAC vocabulary for encoding of the report categories
- To use LGD URIs as location identifiers
- To provide a RDF-based version of data

For instance, the Ushahidi GUI can be extended to let users enter links to relevant LGD entities. These user-identified URIs together with ontology mapping of categories to the MOAC classes make it possible to generate the RDF-based version of data programmatically. Following enrichment with DBpedia data can be automated by wrapping queries presented in *Section 5.3* into the program code.

To the LGD project and DBpedia

Tremendous amount of time in this work was spent while dealing with SPARQL endpoints of LGD and DBpedia. Two main issues, which darkened this experience, included constrains on the number of triples to be retrieved and lacking interoperability of spatial capabilities between different versions of Virtuoso and other systems. Therefore, the recommendations are as follows:

- To provide an endpoint compatible with the GeoSPARQL standard
- To implement better partitioning of the datasets available via APIs

The former recommendation requires additional query processor and SPARQL endpoint built on top of the existing triple store. For instance, Jena and Joseki provide required functionality supporting GeoSPARQL predicates, datatypes and spatial functions.

The latter recommendation would help users to avoid possible frustration provoked by inability to retrieve data of a middle size. Functionality of the endpoints satisfies only those users who want to access small pieces of data or just to explore a few entities. In contrast to this, projects' APIs provide giant subsets of data measured in gigabytes. Manipulations with such data require significant computational efforts. It would be very appreciated if projects provided data portioned, for example, by country.

To the further development of EM ontologies

Two ontologies, the HLX and the MOAC vocabulary, were reviewed in this work. Both of them had their pros and cons. The HLX had better structure and wider representation of EM concepts but lacked terms relevant to Ushahidi categories. The MOAC vocabulary, in contrast, provided needed classes to encode Ushahidi categories but definitely required more structure in the sections responsible for EM concepts.

These observations led to the idea to merge these two vocabularies taking the best from both. For instance, the HLX can be extended with classes and properties to describe report categories.

References

1. Adida, B., & Birbeck, M. (2008). RDFa primer: Bridging the human and data webs. Retrieved June, 20, 2008.
2. Alahmari, F., Thom, J. A., Magee, L., & Wong, W. (2012). Evaluating semantic browsers for consuming linked data. In Proceedings of the Twenty-Third Australasian Database Conference-Volume 124 (pp. 89-98). Australian Computer Society, Inc..
3. Auer, S., Dietzold, S., Lehmann, J., Hellmann, S., & Aumueller, D. (2009). Triplify: light-weight linked data publication from relational databases. In Proceedings of the 18th international conference on World wide web (pp. 621-630). ACM.
4. Babitski G, Bergweiler S, Grebner O, Oberle D, Paulheim H, Probst F (2011) SoKNOS—using semantic technologies in disaster management software. In: The semantic web: research and applications (ESWC 2011), Part II, pp 183–197
5. Ballatore, A., Wilson, D. C., & Bertolotto, M. (2013). A survey of volunteered open geo-knowledge bases in the semantic web. In Quality issues in the management of web information (pp. 93-120). Springer Berlin Heidelberg.
6. Battle, R., Kolas, D. (2012). Enabling the Geospatial Semantic Web with Parliament and GeoSPARQL. Semantic Web (IOS Press) 3(4): 355–370.
7. BBC (28 February 2010). Massive earthquake strikes Chile, BBC News. Retrieved September 1, 2014 from: <http://news.bbc.co.uk/2/hi/8540289.stm>
8. Beckett, D., & McBride, B. (2004). RDF/XML syntax specification (revised). W3C recommendation, 10.
9. Berners-Lee, T. (2006). Linked Data - Design Issues. Retrieved October 1, 2014 from <http://www.w3.org/DesignIssues/LinkedData.html>
10. Berners-Lee, T. (2011). Linked data-design issues (2006). URL <http://www.w3.org/DesignIssues/LinkedData.html>.
11. Berners-Lee, T., Chen, Y., Chilton, L., Connolly, D., Dhanaraj, R., Hollenbach, J., ... & Sheets, D. (2006). Tabulator: Exploring and analyzing linked data on the semantic web. In Proceedings of the 3rd International Semantic Web User Interaction Workshop (Vol. 2006).
12. Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. Scientific American, 284(5), 28-37.
13. Berners-Lee, T., Hollenbach, J., Lu, K., Presbrey, J., & Pru d'ommeaux, E. (2007). Tabulator redux: Writing into the semantic web.
14. Bittner, T., Donnelly, M., & Smith, B. (2009). A spatio-temporal ontology for geographic information integration. International Journal of Geographical Information Science, 23(6), 765-798.
15. Bizer, C., & Gauß, T. (2007). Disco-Hyperdata Browser: A simple browser for navigating the Semantic Web.
16. Bizer, C., & Seaborne, A. (2004). D2RQ-treating non-RDF databases as virtual RDF graphs. In Proceedings of the 3rd international semantic web conference (ISWC2004) (Vol. 2004).
17. Bizer, C., Cyganiak, R., Heath, T. (2007). How to publish Linked Data on the Web. <http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/>
18. Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data-the story so far. International journal on semantic web and information systems, 5(3), 1-22.
19. Borges, M. R., de Faria Cordeiro, K., Campos, M. L. M., & Marino, T. (2011). Linked Open Data and the Design of Information Infrastructure for Emergency Management Systems. In Proceedings of the 10th International ISCRAM Conference, Lisbon, Portugal.

20. Bostock, M., Ogievetsky, V., & Heer, J. (2011). D3: data-driven documents. *Visualization and Computer Graphics*, IEEE Transactions on, 17(12), 2301-2309.
21. Boulos, M. N. K., Resch, B., Crowley, D. N., Breslin, J. G., Sohn, G., Burtner, R., ... & Chuang, K. Y. S. (2011). Crowdsourcing, citizen sensing and sensor web technologies for public and environmental health surveillance and crisis management: trends, OGC standards and application examples. *International journal of health geographics*, 10(1), 67.
22. Bradley, M. (2012). Notes from the Field: Haiti-Displacement and Development in the "Republic of NGOs". Retrieved September 1, 2014 from: <http://www.brookings.edu/blogs/up-front/posts/2012/10/11-haiti-bradley>
23. Breslin J.G, Bojars U, Passant A, Fernandez S, Decker S (2009). SIOC: Content Exchange and Semantic Interoperability Between Social Networks. In: W3C Workshop on the Future of Social Networking. Barcelona, Spain; 2009. Retrieved September 1, 2014 from: <http://www.w3.org/2008/09/msnws/papers/sioc.html>
24. Burke, J. A., Estrin, D., Hansen, M., Parker, A., Ramanathan, N., Reddy, S., & Srivastava, M. B. (2006). Participatory sensing. Center for Embedded Network Sensing.
25. Buscaldi, D., & Rosso, P. (2008). Geo-WordNet: Automatic Georeferencing of WordNet. In LREC.
26. Chrisman, N. (2001). *Exploring geographic information systems*. Wiley.
27. Clark, L. (2010, November). Sparql views: A visual sparql query builder for drupal. In 9th International Semantic Web Conference, ISWC.
28. Cox, S., & Schade, S. (2010, May). Linked data: What does it offer earth sciences?. In EGU General Assembly Conference Abstracts (Vol. 12, p. 2079).
29. Dadzie, A., & Rowe, M. (2011) Approaches to Visualising Linked Data : A Survey. *Semantic Web Journal*, 1(2), 34. doi:10.3233/SW-2011-0037
30. Datalift (2014). Datalift promotes the Web of Data. Retrieved September 1, 2014 from Datalift: <http://datalift.org/>
31. de Faria Cordeiro, K., Marino, T., Campos, M. L. M., & Borges, M. (2011, June). Use of Linked Data in the design of information infrastructure for collaborative emergency management system. In *Computer Supported Cooperative Work in Design (CSCWD)*, 2011 15th International Conference on(pp. 764-771). IEEE.
32. de León, A., Saquicela, V., Vilches, L. M., Villazón-Terrazas, B., Priyatna, F., & Corcho, O. (2010, September). Geographical linked data: a Spanish use case. In *Proceedings of the 6th International Conference on Semantic Systems* (p. 36). ACM.
33. Dean, M., Schreiber, G., Bechhofer, S., van Harmelen, F., Hendler, J., Horrocks, I., ... & Stein, L. A. (2004). OWL web ontology language reference. W3C Recommendation February, 10. Retrieved September 1, 2014 from: <http://www.w3.org/TR/owl-ref/>
34. Deligiannidis, L., Kochut, K. J., & Sheth, A. P. (2007, November). RDF data exploration and visualization. In *Proceedings of the ACM first workshop on CyberInfrastructure: information management in eScience* (pp. 39-46). ACM.
35. Dietzold, S (2014). DBpedia about and news. Retrieved September 1, 2014 from DBpedia: <http://dbpedia.org/About>
36. Duc, K. N., Vu, T. T., & Ban, Y. (2014). Ushahidi and Sahana Eden Open-Source Platforms to Assist Disaster Relief: Geospatial Components and Capabilities. In *Geoinformation for Informed Decisions* (pp. 163-174). Springer International Publishing.
37. Egenhofer, M. J. (2002, November). Toward the semantic geospatial web. In *Proceedings of the 10th ACM international symposium on Advances in geographic information systems* (pp. 1-4). ACM.

38. Feliachi, A., Abadie, N., Hamdi, F., & Atemezing, G. A. (2013). Interlinking and visualizing linked open data with geospatial reference data. In OM (pp. 237-238).
39. Fellbaum, C. (1998.): WordNet: An electronic lexical database. MIT press, Cambridge, MA
40. Frasinca, F., Telea, A., & Houben, G. J. (2006). Adapting graph visualization techniques for the visualization of RDF data. In Visualizing the semantic web (pp. 154-171). Springer London.
41. GeoNames (n.d.). GeoNames about page. Retrieved September 1, 2014 from GeoNames: <http://www.geonames.org/about.html>
42. Giunchiglia, F., Maltese, V., Farazi, F., & Dutta, B. (2010). GeoWordNet: a resource for geospatial applications. In The Semantic Web: Research and Applications (pp. 121-136). Springer Berlin Heidelberg.
43. Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4), 211-221.
44. Goodchild, M. F., & Glennon, J. A. (2010). Crowdsourcing geographic information for disaster response: a research frontier. *International Journal of Digital Earth*, 3(3), 231-241.
45. Goodwin, J., Dolbear, C., & Hart, G. (2008). Geographical linked data: The administrative geography of Great Britain on the semantic web. *Transactions in GIS*, 12(s1), 19-30.
46. Goyal, S., & Westenthaler, R. (2004). Rdf gravity (rdf graph visualization tool). Salzburg Research, Austria.
47. Graves, A. (2013, June). Creation of visualizations based on linked data. In Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics (p. 41). ACM.
48. Gruber, T. (2007). Ontology of folksonomy: A mash-up of apples and oranges. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 3(1), 1-11.
49. Guarino, N. (Ed.). (1998). Formal ontology in information systems: Proceedings of the first international conference (FOIS'98), June 6-8, Trento, Italy (Vol. 46). IOS press.
50. Haggarty, A., Naidoo, S. (2008): Global Symposium+5 Final Report, Information for Humanitarian Action. UN OCHA, Geneva.
51. Haklay, M., Singleton, A., & Parker, C. (2008). Web mapping 2.0: The neogeography of the GeoWeb. *Geography Compass*, 2(6), 2011-2039.
52. Hattotuwa, S., & Stauffacher, D. (2011). Haiti and beyond: getting it Right in Crisis Information Management. In Peacebuilding in the Information Age: Sifting hype from reality (pp. 9-11). ICT4Peace Foundation; United States. Harvard University. Berkman Centre for Internet and Society; Georgia Institute of Technology (GeorgiaTech).
53. Heim, P., Thom, D., & Ertl, T. (2011). SemSor: combining social and semantic web to support the analysis of emergency situations. In Proceedings of the 2nd Workshop on Semantic Models for Adaptive Interactive Systems SEMAIS.
54. Hoffart, J., Suchanek, F. M., Berberich, K., & Weikum, G. (2013). YAGO2: a spatially and temporally enhanced knowledge base from Wikipedia. *Artificial Intelligence*, 194, 28-61.
55. Howe, J. (2006). The rise of crowdsourcing. *Wired magazine*, 14(6), 1-4.
56. IASC (2013). About The Inter-Agency Standing Committee. Retrieved September 1, 2014 from IASC: <http://www.humanitarianinfo.org/iasc/>
57. Inter-Agency Standing Committee (2006). Guidance note on using the cluster approach to strengthen humanitarian response. In Guidance note on using the cluster approach to strengthen humanitarian response. IASC.
58. Janowicz, K., & Hitzler, P. (2012). The digital earth as knowledge engine. *Semantic Web*, 3(3), 213-221.
59. Janowicz, K., Scheider, S., Pehle, T., & Hart, G. (2012). Geospatial semantics and linked spatiotemporal data—Past, present, and future. *Semantic Web*, 3(4), 321-332.

60. Keßler, C. & Hendrix, C. (forthcoming) The Humanitarian eXchange Language: Coordinating Disaster Response with Semantic Web Technologies. *Semantic Web Journal*, accepted.
61. Klyne, G., & Carroll, J. J. (2005). Resource Description Framework (RDF): Concepts and Abstract Syntax. W3C Recommendation.
62. Kolas, D., Emmons, I., & Dean, M. (2009). Efficient linked-list rdf indexing in parliament. *SSWS*, 9, 17-32.
63. Kyzirakos K., M. Karpathiotakis, M. Koubarakis (2012). Strabon: A Semantic Geospatial DBMS. In ISWC, pp. 295-311, 2012
64. Larsen, L. (2007). Strengthening Humanitarian Information Management: A Status Report. Report, Field Information Services Unit, United Nations Office for Coordination of Humanitarian Affairs, Geneva.
65. Lassila, O. (2006) Browsing the Semantic Web. 17th International Conference on Database and Expert Systems Applications (DEXA'06), 5th International Workshop on Web Semantics, pp.365-369, Krakow (Poland), September 2006.
66. Lemmens R, Kessler C (2014) Geo-information visualizations of linked data. In: Proceedings of the 17th AGILE conference on geographic information science, 3–6 June 2014, Castelln, Spain
67. Lieberman, J., Singh, R., Goad, C. (2007). W3C Geospatial Ontologies. W3C Incubator Group. Retrieved from <http://www.w3.org/2005/Incubator/geo/XGR-geo-ont-20071023/>
68. Limbu, M. (2012). Management of a Crisis (MOAC) Vocabulary Specification. Retrieved September 1, 2014 from: <http://observedchange.com/moac/ns/>
69. LinkingOpenData (2014). The W3C SWEOW Linking Open Data community project. In W3C WIKI. Retrieved September 1, 2014 from the W3C WIKI: <http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData>
70. Liu, S., Shaw, D., & Brewster, C. (2013, May). Ontologies for crisis management: a review of state of the art in ontology design and usability. In Proceedings of the Information Systems for Crisis Response and Management conference (ISCRAM 2013 12-15 May, 2013).
71. Liu, S., Shaw, D., & Brewster, C. (2013, May). Ontologies for crisis management: a review of state of the art in ontology design and usability. In Proceedings of the Information Systems for Crisis Response and Management conference (ISCRAM 2013 12-15 May, 2013).
72. LOV (2014). In LOV at a glanceRetrieved September 1, 2014 from LOV: <http://lov.okfn.org/dataset/lov/about/>
73. Mazzocchi, S. & Ciccamese, P., (2008). About Welkin. SIMILE | Welkin. Retrieved September 1, 2014 from WELKIN: <http://simile.mit.edu/welkin/>
74. McDonald, B., & Gordon, P. (2008). United Nations' Efforts to Strengthen Information Management for Disaster Preparedness and Response. *NATURAL*, 59.
75. Mendes, P. N., Jakob, M., García-Silva, A., & Bizer, C. (2011, September). DBpedia spotlight: shedding light on the web of documents. In Proceedings of the 7th International Conference on Semantic Systems (pp. 1-8). ACM.
76. Mijovic, V., Janev, V., & Vrane, S. (2013, August). Main Challenges in Using LOD in Emergency Management. In 2012 23rd International Workshop on Database and Expert Systems Applications (pp. 21-25). IEEE.
77. Milis, K., & Van de Walle, B. (2007). IT for corporate crisis management: findings from a survey in 6 different industries on management attention, intention and actual use. In System Sciences, 2007. HICSS 2007. 40th Annual Hawaii International Conference on (pp. 24-24). IEEE.
78. Morrow, N., Mock, N., Papendieck, A., & Kocmich, N. (2011). Independent evaluation of the Ushahidi Haiti project. *Development Information Systems International*, 8, 2011.
79. Munro, R. (2013). Crowdsourcing and the crisis-affected community. *Information retrieval*, 16(2), 210-266.

80. Okolloh, O. (2008) Ushahidi, or 'testimony': Web 2.0 tools for crowdsourcing crisis information. *Participatory Learning and Action* 59, 65-70.
81. O'reilly, T. (2007). What is Web 2.0: Design patterns and business models for the next generation of software. *Communications and Strategies*, 65(1), 17-37.
82. OpenStreetMap (2014). OpenStreetMap stats report. Retrieved October 1, 2014 from http://www.openstreetmap.org/stats/data_stats.html
83. Ortmann, J., Limbu, M., Wang, D., & Kauppinen, T. (2011). Crowdsourcing linked open data for disaster management. *Terra Cognita*, 11-22.
84. Patroumpas, K., Alexakis, M., Giannopoulos, G., & Athanasiou, S. (2014a). TripleGeo: an ETL Tool for Transforming Geospatial Data into RDF Triples. In *EDBT/ICDT Workshops* (pp. 275-278).
85. Patroumpas, K., Giannopoulos, G., & Athanasiou, S. (2014b) Towards GeoSpatial Semantic Data Management: Strengths, Weaknesses, and Challenges Ahead. <http://www.dbnet.ece.ntua.gr/pubs/uploads/TR-2014-13.pdf>
86. Poblet, M., García-Cuesta, E., & Casanovas, P. (2014, January). IT Enabled Crowds: Leveraging the Geomobile Revolution for Disaster Management. In *Sintelnet WG5 Workshop on Crowd Intelligence: Foundations, Methods and Practices*.
87. Prud'hommeaux, E., & Lee, R. (2004). W3C RDF validation service. Retrieved September 1, 2014 from <http://www.w3.org/RDF/Validator>.
88. Rusher, J. (2003). Triple store. In *Workshop on Semantic Web Storage and Retrieval-Position Paper*.
89. Rutledge, L., Van Ossenbruggen, J., & Hardman, L. (2005, May). Making RDF presentable: integrated global and local semantic Web browsing. In *Proceedings of the 14th international conference on World Wide Web* (pp. 199-206). ACM.
90. Scharffe, F., Ateazing, G., Troncy, R., Gandon, F., Villata, S., Bucher, B., ... & Vatan, B. (2012, July). Enabling linkeddata publication with the datalift platform. In *Proc. AAI workshop on semantic cities*.
91. Schulz, A., Paulheim, H., & Probst, F. (2012, April). Crisis information management in the web 3.0 age. In *Proceedings of the Information Systems for Crisis Response and Management Conference (ISCRAM 2012)* (pp. 1-6).
92. Schwering, A. (2008). Approaches to Semantic Similarity Measurement for Geo-Spatial Data: A Survey. *Transactions in GIS*, 12(1), 5-29.
93. Shvaiko, P., Farazi, F., Maltese, V., Ivanyukovich, A., Rizzi, V., Ferrari, D., & Ucelli, G. (2012). Trentino government linked open geo-data: a case study. In *The Semantic Web-ISWC 2012* (pp. 196-211). Springer Berlin Heidelberg.
94. Skjæveland, M. (2012). Sgvizler: A javascript wrapper for easy visualization of sparql result sets. In *Extended Semantic Web Conference*.
95. Sabol, V., Tschinkel, G., Veas, E., Hoefler, P., Mutlu, B., & Granitzer, M. (2014). Discovery and visual analysis of linked data for humans. In *The Semantic Web-ISWC 2014* (pp. 309-324). Springer International Publishing.
96. Suchanek, F. M., Kasneci, G., & Weikum, G. (2008). Yago: A large ontology from wikipedia and wordnet. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(3), 203-217.
97. Terpstra, T., de Vries, A., Stronkman, R., & Paradies, G. L. (2012, April). Towards a realtime Twitter analysis during crises for operational crisis management. In *ISCRAM'12: Proceedings of the 9th International ISCRAM Conference*.
98. The Dublin Core Metadata Initiative (2014). Background. Retrieved September 1, 2014 from: <http://dublincore.org/metadata-basics/>

99. The Standby Task Force (2014). About Standby Task Force. Retrieved September 1, 2014 from: <http://blog.standbytaskforce.com/about-2/>
100. Tramp, S., Van Nuffelen, B., Frischmuth, P., & Auer, S. (2011) Creating Knowledge out of Interlinked Data: The Integrated LOD2 Tool Stack.
101. TripleGeo (2014). TripleGeo utility for converting geospatial data into triples. Retrieved September 1, 2014 from the TripleGeo GitHub: <https://github.com/GeoKnow/TripleGeo>
102. Turner, A. (2006). Introduction to neogeography. " O'Reilly Media, Inc."
103. United States Geological Survey (1 March 2010). Magnitude 8.8 - OFFSHORE BIO-BIO, CHILE. USGS Earthquake Summary. Retrieved September 1, 2014 from: <http://earthquake.usgs.gov/earthquakes/eqinthenews/2010/us2010tfan/#summary>.
104. UNOCHA (2006). Guidelines for OCHA Field Information Management. Office for the Coordination of Humanitarian Affairs (OCHA), United Nations, New York, United Nations.
105. Volz, J., Bizer, C., Gaedke, M., & Kobilarov, G. (2009). Silk-A Link Discovery Framework for the Web of Data. LDOW, 538.
106. W3C (2012). Examples of RDF Validation. Retrieved September 1, 2014 from W3C: <http://www.w3.org/2012/12/rdf-val/SOTA>
107. W3C (2014a). Vocabularies. Retrieved September 1, 2014 from: <http://www.w3.org/standards/semanticweb/ontology>.
108. W3C (2014b). What is Query Used For? Retrieved September 1, 2014 from: <http://www.w3.org/standards/semanticweb/query.html>
109. W3C Semantic Web Interest Group (2006). W3C Semantic Web Interest Group: Basic Geo (WGS84 lat/long) Vocabulary. Retrieved September 1, 2014 from: <http://www.w3.org/2003/01/geo/>
110. Walle, B. V. D., & Dugdale, J. (2012). Information management and humanitarian relief coordination: findings from the Haiti earthquake response. *International Journal of Business Continuity and Risk Management*, 3(4), 278-305.
111. Wikimapia (2014). Wikimapia statistic page. Retrieved October 1, 2014 from <http://wikimapia.org/>
112. Zook, M., Graham, M., Shelton, T., & Gorman, S. (2010). Volunteered geographic information and crowdsourcing disaster relief: a case study of the Haitian earthquake. *World Medical & Health Policy*, 2(2), 7-33.
113. Zviedris, M., & Barzdins, G. (2011). ViziQuer: a tool to explore and query SPARQL endpoints. In *The Semantic Web: Research and Applications* (pp. 441-445). Springer Berlin Heidelberg.

Appendices

Appendix A. Example of Ushahidi data

This appendix provides two example reports (4773, 4809) downloaded from the Ushahidi API. First, the reports are presented in the original JSON format. The following table gives a tabular representation of the same reports.

In JSON

```

1 {
2   "payload": {
3     "incidents": [
4       {
5         "incident": {
6           "incidentid": "4773",
7           "incidenttitle": "Se establecio un hospital mobil en Angol / Established hospital in Angol",
8           "incidentdescription": "Se establecio un hospital mobil en Angol, Region Araucania.\n\n"-
9             The USAID/OFDA-funded U.S. Air Force Expeditionary Medical Support (EMEDS) is fully operational-
10            with doctors seeing their first patients on March 13. The EMEDS unit is located in Angol town, La Araucania Region..\n",
11           "incidentdate": "2010-03-15 20:29:00",
12           "incidentmode": "1",
13           "incidentactive": "1",
14           "incidentverified": "0",
15           "locationid": "23933",
16           "locationname": "(-37.803002585189645, -72.7016830444336)",
17           "locationlatitude": "-37.803388",
18           "locationlongitude": "-72.700921"
19         },
20       },
21     ],
22     {
23       "incident": {
24         "incidentid": "4809",
25         "incidenttitle": "Protesta impide transito (SMS 9683XXXX)",
26         "incidentdescription": "Protesta cerca del puerto de talcahuano impide
27           transito desde y hacia conce. Protest impedes traffic on Talcahuano bridge.",
28         "incidentdate": "2010-03-18 11:47:00",
29         "incidentmode": "1",
30         "incidentactive": "1",
31         "incidentverified": "0",
32         "locationid": "23972",
33         "locationname": "puerto de talcahuano",
34         "locationlatitude": "-36.767247",
35         "locationlongitude": "-73.087776"
36       },
37     },
38   ],
39 }

```

In tabular format

Serial number	INC.TITLE	INCIDENT DATE	LOCATION	DESCRIPTION	CATEGORY	LAT	LONG	APPROVED	VERIFIED
4809	Protesta impide transito (SMS 9683XXXX)	2010-03-18	puerto de talcahuano	Protesta cerca del puerto de talcahuano impide transito desde y hacia conce. Protest impedes traffic on Talcahuano bridge.	3.Catastro	- 36.76 725	- 73.087 776	yes	no
4773	Se establecio un hospital mobil en Angol / Established hospital in Angol	2010-03-15	- 37.80300258 5189645, - 72.70168304 44336	Se establecio un hospital mobil en Angol, Region Araucania. \"- The USAID/OFDA-funded U.S. Air Force Expeditionary Medical Support (EMEDS) is fully operational—with doctors seeing their first patients on March 13. The EMEDS unit is located in Angol town, La Araucania Region..."	4a. Servicios de Salud	- 37.80 339	- 72.700 921	yes	no

Appendix B. Ontology mapping between the Ushahidi categories and MOAC for Chile

The following table provides an ontology mapping between the Ushahidi categories used in Chile and the classes of the Management of a Crisis vocabulary. This ontology mapping helped to represent the semantics of the Ushahidi reports allowing multi criteria filtering of the reports.

Ushahidi categories used during Chilean Earthquake 2010			MOAC terms Prefix MOAC: < http://observedchange.com/moac/ns/# >
Category number	Spanish	English translation	
1	Emergencia	Emergency	MOAC:Emergency
1a	Estructura Colapsada	Collapsed structure	MOAC:CollapsedStructure
1b	Incendio	Fire	MOAC:Fire
1c	Personas atrapadas	Trapped people	MOAC:PeopleTrapped
1d	Urgencias Medicas	Medical emergency	MOAC:MedicalEmergency
1e	Tsunami	Tsunami	
1f	Replicas	Aftershock	MOAC:EarthquakeAndAftershock
2	Amenazas	Menace	MOAC:Menaces
2a	Estructuras en Riesgo	Unstable Structure	MOAC:UnstableStructure
2b	Saqueos	Looting	MOAC:Looting
2c	Problemas de Seguridad	Problems with security	MOAC:SecurityConcern
3	Catastro	Damaged infrastructure	MOAC:InfrastructureDamage
3a	Desabastecimiento de Agua	Water shortage	MOAC:WaterShortage
3b	Ruta Bloqueada	Blocked Road	MOAC:RoadBlocked
3c	Cortes de Electricidad	Power outage	MOAC:PowerOutage
3d	Desabastecimiento de Alimentos	Food shortage	MOAC:FoodShortage
3e	Desabastecimiento de Medicamentos	Drug shortage	MOAC:MedicalEquipmentAndSupplyNeeds
3g	Viviendas afectadas	Affected households	MOAC:AffectedPopulation
3i	Familias Afectadas	Affected families	MOAC:AffectedPopulation
3f	Desabastecimiento de Combustible	Fuel shortage	MOAC:FuelShortage
4	Respuesta	Response	MOAC:ServiceAvailable
4a	Servicios de Salud	Health service	MOAC:HospitalOperating
4b	Búsqueda y Rescate	Search and rescue	MOAC:SearchAndRescue
4c	Refugio Albergue	Shelter	MOAC:ShelterOffered
4d	Desabastecimiento de Alimentos	Distribution of food	MOAC:FoodDistributionPoint
4e	Saneamiento de Agua	Clean water	MOAC:WaterSanitationAndHygienePromotion
4f	Recepción de Ayuda	Aid distribution	MOAC:NonfoodAidDistributionPoint
4h	Morgue	Morgue	MOAC:HumanRemainsManagement
4i	Distribución de Agua	Distribution of water	MOAC:WaterDistributionPoint
4j	Comisarias y Carabineros	Police and Military forces	
4k	Servicios Telefonicos	Telephone service	
5	Noticias de Personas	Information about people	MOAC:PersonsNews
5a	Decesos	Deaths	MOAC:Deaths

Ushahidi categories used during Chilean Earthquake 2010			MOAC terms
Category number	Spanish	English translation	Prefix MOAC: <http://observedchange.com/moac/ns/#>
5b	Personas Desaparecidas	Missing people	MOAC:MissingPersons
5c	Peticiones de envios de mensajes	Request to forward a message	MOAC:AskingToFowardAMessage
6	Comercio Abierto	Available trade	
6a	Farmacias	Pharmacy	
6b	Supermercado	Supermarket	
6c	Bencineras	Gas station	
7	Locacion sin Ayuda	Locations where humanitarian response is needed	MOAC:NeedsResponse3W
8	Donaciones	Donations	
8a	Donaciones de Sangre	Blood donation	
8b	Donaciones de Dinero	Money donation	MOAC:FinancialServicesAvailable
8c	Donacion de Especies	Other donations	
9	Voluntarios	Volunteers	
9a	Voluntarios de Salud	Medical Volunteers	
9b	Voluntarios en Ayuda Humanitaria	Humanitarian Aid Volunteers	
9c	Voluntarios en Vivienda y Construccion	Volunteers for assistance in construction works	

Appendix C. Ontology mapping between the Ushahidi categories and MOAC for Haiti

This table is given to show the similarities between different systems of the Ushahidi categories used in Haiti and Chile. In addition, the relevant MOAC classes are provided to emphasize the general applicability of the MOAC vocabulary for the representation of the Ushahidi categories.

Haitian Earthquake 2010		Chilean Earthquake 2010				
Category number	English name	Category number	Spanish	English translation	MOAC terms Prefix MOAC: < http://observedchange.com/moac/ns/# >	Terms from other vocabularies
1	Emergency	1	Emergencia	Emergency	MOAC:Emergency	
5a	Collapsed structure,	1a	Estructura Colapsada	Collapsed structure	MOAC:CollapsedStructure	
		1b	Incendio	Fire	MOAC:Fire	
1c	Trapped people	1c	Personas atrapadas	Trapped people	MOAC:PeopleTrapped	
1b	Medical Emergency	1d	Urgencias Medicas	Medical emergency	MOAC:MedicalEmergency	
		1e	Tsunami	Tsunami		http://ontology.es/WordNet/data/Tsunami
6c	Earthquake and aftershocks	1f	Replicas	Aftershock	MOAC:EarthquakeAndAftershock	
4	Security Threats	2	Amenazas	Menace	MOAC:Menaces	
5b	Unstable Structure	2a	Estructuras en Riesgo	Unstable Structure	MOAC:UnstableStructure	
4a	Looting	2b	Saqueos	Looting	MOAC:Looting	
		2c	Problemas de Seguridad	Problems with security	MOAC:SecurityConcern	
5	Infrastructure Damage	3	Catastro	Damaged infrastructure	MOAC:InfrastructureDamage	
2b	Water shortage	3a	Desabastecimiento de Agua	Water shortage	MOAC:WaterShortage	
5c	Road blocked	3b	Ruta Bloqueada	Blocked Road	MOAC:RoadBlocked	
		3c	Cortes de Electricidad	Power outage	MOAC:PowerOutage	
2a	Food Shortage	3d	Desabastecimiento de Alimentos	Food shortage	MOAC:FoodShortage	
3c	Medical equipment	3e	Desabastecimiento de	Drug shortage	MOAC:MedicalEquipmentAndSupplyNeeds	

	and supply needs		Medicamentos			
		3g	Viviendas afectadas	Affected households	MOAC:AffectedPopulation	
		3i	Familias Afectadas	Affected families	MOAC:AffectedPopulation	
		3f	Desabastecimiento de Combustible	Fuel shortage	MOAC:FuelShortage	
7	Services Available	4	Respuesta	Response	MOAC:ServiceAvailable	
7d	Hospital/Clinics Operating	4a	Servicios de Salud	Health service	MOAC:HospitalOperating	
		4b	Búsqueda y Rescate	Search and rescue	MOAC:SearchAndRescue	
		4c	Refugio Albergue	Shelter	MOAC:ShelterOffered	
7a	Food distribution point	4d	Desabastecimiento de Alimentos	Distribution of food	MOAC:FoodDistributionPoint	
		4e	Saneamiento de Agua	Clean water	MOAC:WaterSanitationAndHygienePromotion	
7c	Non-food aid distribution point	4f	Recepción de Ayuda	Aid distribution	MOAC:NonfoodAidDistributionPoint	
		4h	Morgue	Morgue	MOAC:HumanRemainsManagement	
7b	Water distribution point	4i	Distribución de Agua	Distribution of water	MOAC:WaterDistributionPoint	
		4j	Comisarias y Carabineros	Police and Military forces		
		4k	Servicios Telefonicos	Telephone service		http://139.91.183.30:9090/RDF/VRP/Examples/DCD100.rdf#TelephoneService
		5	Noticias de Personas	Information about people	MOAC:PersonsNews	
6a	Deaths	5a	Decesos	Deaths	MOAC:Deaths	
6b	Missing people	5b	Personas Desaparecidas	Missing people	MOAC:MissingPersons	
		5c	Peticiones de envios de mensajes	Request to forward a message	MOAC:AskingToForwardAMessage	

		6	Comercio Abierto	Available trade		http://schema.org/LocalBusiness
		6a	Farmacias	Pharmacy		http://schema.org/Pharmacy
		6b	Supermercado	Supermarket		http://schema.org/Store
		6c	Bencineras	Gas station		http://www.losa-cnr.it/ontologies/WordNet/OWN#GASOLINE_STATION_GAS_STATION_FILLING_STATION_PETROL_STATION
		7	Locacion sin Ayuda	Locations where humanitarian response is needed	MOAC:NeedsResponse3W	
		8	Donaciones	Donations		http://truesense.net/wordnet.01.owl#donation
		8a	Donaciones de Sangre	Blood donation		
		8b	Donaciones de Dinero	Money donation	MOAC:FinancialServicesAvailable	
		8c	Donacion de Especies	Other donations		
		9	Voluntarios	Volunteers		http://dbpedia.org/class/yago/Volunteer10759151
		9a	Voluntarios de Salud	Medical Volunteers		
		9b	Voluntarios en Ayuda Humanitaria	Humanitarian Aid Volunteers		
		9c	Voluntarios en Vivienda y Construccion	Volunteers for assistance in construction works		

Appendix D. List of missing LGD objects

List of objects that were not exposed via the SPARQL endpoint of the LGD project. They were downloaded using the LGD API.

1. <http://linkedgedata.org/triplify/way111133067>
2. <http://linkedgedata.org/triplify/way113992956>
3. <http://linkedgedata.org/triplify/way115771187>
4. <http://linkedgedata.org/triplify/way122150522>
5. <http://linkedgedata.org/triplify/way136999265>
6. <http://linkedgedata.org/triplify/way138296132>
7. <http://linkedgedata.org/triplify/way139960201>
8. <http://linkedgedata.org/triplify/way141025604>
9. <http://linkedgedata.org/triplify/way144276507>
10. <http://linkedgedata.org/triplify/way162155089>
11. <http://linkedgedata.org/triplify/way165561550>
12. <http://linkedgedata.org/triplify/way197582048>
13. <http://linkedgedata.org/triplify/way207631268>
14. <http://linkedgedata.org/triplify/way22892350>
15. <http://linkedgedata.org/triplify/way238811346>
16. <http://linkedgedata.org/triplify/way25422270>
17. <http://linkedgedata.org/triplify/way25422319>
18. <http://linkedgedata.org/triplify/way27240910>
19. <http://linkedgedata.org/triplify/way27726459>
20. <http://linkedgedata.org/triplify/way27729607>
21. <http://linkedgedata.org/triplify/way39487867>
22. <http://linkedgedata.org/triplify/way48592279>
23. <http://linkedgedata.org/triplify/way48592362>
24. <http://linkedgedata.org/triplify/way48592741>
25. <http://linkedgedata.org/triplify/way49478550>
26. <http://linkedgedata.org/triplify/way51464187>
27. <http://linkedgedata.org/triplify/way51483976>
28. <http://linkedgedata.org/triplify/way51483977>
29. <http://linkedgedata.org/triplify/way51887764>
30. <http://linkedgedata.org/triplify/way52955215>
31. <http://linkedgedata.org/triplify/way87098027>

Appendix E. Table of namespace prefixes

The table below summarizes namespace prefixes used in the queries given in this thesis. The first column introduces the prefixes, the second column provides the full URIs of the prefixes, and the last column gives a short description of the resources.

Namespace prefix	Namespace IRI	Resource description
rdf	< http://www.w3.org/1999/02/22-rdf-syntax-ns# >	Vocabulary defined by the RDF specification
rdfs	< http://www.w3.org/2000/01/rdf-schema# >	Resource Description Framework Schema vocabulary, an extension built on the limited vocabulary of RDF
xsd	< http://www.w3.org/2001/XMLSchema# >	Vocabulary for datatypes implemented by W3C XML Schema Definition Language
dc11	< http://purl.org/dc/elements/1.1/ >	The Dublin Core Metadata Element Set, consisting of 15 basic terms. Published as IETF RFC 5013, ANSI/NISO Standard Z39.85-2007, and ISO Standard 15836:2009
dcterms	< http://purl.org/dc/terms/ >	The Dublin Core Metadata Initiative (DCMI) Metadata Terms is an extended vocabulary built on the Dublin Core Metadata Element Set
MOAC	< http://observedchange.com/moac/ns/# >	Management of a Crisis (MOAC) Vocabulary
dbo	< http://dbpedia.org/ontology/ >	DBpedia ontology
dbr	< http://dbpedia.org/resource/ >	Namespace of DBpedia resources
dbprop	< http://dbpedia.org/property/ >	Set of properties (predicates) used for DBpedia resources
lgdr	< http://linkedgeo.org/triplify/ >	Namespace of LinkedGeoData resources
lgdo	< http://linkedgeo.org/ontology/ >	LinkedGeoData ontology
sf	< http://www.opengis.net/ont/sf# >	OGC Simple Features ontology
units	< http://www.opengis.net/def/uom/OGC/1.0/ >	OGC Units of Measure 1.0 vocabulary
w3geo	< http://www.w3.org/2003/01/geo/wgs84_pos# >	W3C Basic Geo (WGS84 lat/long) vocabulary
geo	< http://www.opengis.net/ont/GeoSPARQL# >	OGC GeoSPARQL ontology
geom	< http://geovocab.org/geometry# >	NeoGeo Geometry Ontology. A vocabulary for describing geographical regions in RDF.
geof	< http://www.opengis.net/def/function/GeoSPARQL/ >	Topological functions described in the GeoSPARQL standard
geos	< http://geovocab.org/spatial# >	NeoGeo Spatial Ontology. A vocabulary for describing topological relations between features.

Appendix F. List of the selected reports.

This appendix gives the list of the selected reports. These reports were taken from the original data set for further establishing of semantic links. Only these reports were used for the construction of the proof of concept.

Number	Report ID	Number	Report ID
1	4785	41	4559
2	4308	42	3963
3	4371	43	4347
4	4572	44	3724
5	4376	45	4346
6	4172	46	4135
7	4773	47	4345
8	4305	48	4686
9	4106	49	4122
10	4702	50	4136
11	3704	51	4286
12	4104	52	4280
13	4398	53	4211
14	4103	54	3803
15	3904	55	3968
16	4720	56	4210
17	4056	57	3756
18	4518	58	4125
19	4566	59	3735
20	4672	60	4282
21	4102	61	4349
22	3982	62	4730
23	4517	63	3726
24	3959	64	4004
25	4673	65	4229
26	4268	66	4960
27	4441	67	4214
28	4655	68	4213
29	4255	69	4002
30	4535	70	4212
31	4276	71	3920
32	4266	72	4010
33	4267	73	4344
34	3925	74	3960
35	4902	75	4343
36	4568	76	4849
37	4809	77	4001
38	4598	78	4352
39	4567	79	4301
40	4348	80	4119
		81	3962