

# GI Systems for Public Health with an Ontology Based Approach

Master Thesis for Geospatial Technologies



International Erasmus Mundus Master. Program with cooperation of:

*University of Münster (WWU), Institute for Geoinformatics (ifgi) - Germany;  
Universidade Nova de Lisboa (UNL), Instituto Superior de Estatística e Gestão de  
Informação (ISEGI), Lisboa, Portugal;  
Universitat Jaume I (UJI), Institute of New Imaging Technologies (INIT), Castellón,  
Spain.*

**prepared by  
Nurefşan Gür**

Castellón de la Plana, February 2012

## **Foreword/Preface**

I hereby, certify that I have written this thesis independently and with no other tools than the specified. Data is provided by CSISP-Valencia with the given right of implementing and publishing in Linked Open Data. Sources used are indicated as secondary literature and referenced in the bibliography. Final work is subject to submission by approval of three supervisors as listed in dissertation board.

Castellon, February 2012

## **Dissertation Board**

Laura Diaz Sanchez, PhD, Postdoctoral researcher -INIT, Institute of New Imaging Technologies, Universitat Jaume I – Spain

Tomi Kauppinen, PhD, Postdoctoral researcher - IFGI, Institute for Geoinformatics University of Münster – Germany

Prof. Marco Painho - ISEGI, Instituto Superior de Estatística e Gestão de Informação Universidade NOVA de Lisboa – Portugal

## *Acknowledgements*

*This study wouldn't be complete without the support of my supervisors. I would like to express my gratitude for their help and guidance.*

*I would like thank my supervisor Laura Diaz for countless times reading my drafts and improving the ideas.*

*And thanks to my co-supervisor Tomi Kauppinen for the constructive discussions and leading to valuable sources. I*

*also would like to thank my second co-supervisor Prof.*

*Marco Painho for his contributive feedback to the work.*

*I extend my thanks: To my family for their patience, to my friends for their support and those for being my family...*

# **GI Systems for Public Health with an Ontology Based Approach**

## **Abstract**

Health is an indispensable attribute of human life. In modern age, utilizing technologies for health is one of the emergent concepts in several applied fields. Computer science, (geographic) information systems are some of the interdisciplinary fields which motivates this thesis.

Inspiring idea of the study is originated from a rhetorical disease DbHd: Database Hugging Disorder, defined by Hans Rosling at World Bank Open Data speech in May 2010. The cure of this disease can be offered as linked open data, which contains ontologies for health science, diseases, genes, drugs, GEO species etc. LOD-Linked Open Data provides the systematic application of information by publishing and connecting structured data on the Web.

In the context of this study we aimed to reduce boundaries between semantic web and geo web. For this reason a use case data is studied from Valencia CSISP- Research Center of Public Health in which the mortality rates for particular diseases are represented spatio-temporally. Use case data is divided into three conceptual domains (health, spatial, statistical), enhanced with semantic relations and descriptions by following Linked Data Principles. Finally in order to convey complex health-related information, we offer an infrastructure integrating geo web and semantic web. Based on the established outcome, user access methods are introduced and future researches/studies are outlined.

## **Keywords**

Linked Open Data

Semantic Web

Health Statistics

Geographical Information Systems

Spatial / Health / Statistical Ontologies

## Acronyms

**URI** – Unified Resource Identifier

**URL** – Unified Resource Locator

**RDF** – Resource Description Framework

**HTTP** – Hypertext Transfer Protocol

**WWW** – World Wide Web

**XML** – Extensible Markup Language

**SPARQL** – SPARQL Protocol and RDF Query Language

**RDB** – Relational Database

**RDBMS** – Relational Database Management System

**CSV** – Comma Separated Values

**LOD** – Linked Open Data

**OWL** – Ontology Web Language

**GIS** – Geographic Information Systems

**GIScience** – Geographic Information Science

**CSISP** – Centro Superior de Investigación en Salud Pública

**W<sub>3</sub>C** – World Wide Web Consortium

**FOAF** – Friend of a friend

**SIOC** – Semantically interlinked online communities

**CKAN** – Comprehensive Knowledge Archive Network

**HCLSIG** – Health Care and Life Sciences Interest Group

# Contents

<b>Dissertation Board .....</b>	<b>ii</b>
<b>Acknowledgements .....</b>	<b>iii</b>
<b>Abstract .....</b>	<b>iv</b>
<b>Keywords .....</b>	<b>v</b>
<b>Acronyms .....</b>	<b>vi</b>
<b>Index of Figures .....</b>	<b>ix</b>
<b>Index of Tables .....</b>	<b>ix</b>
<b>Index of Listings .....</b>	<b>ix</b>
<b>CHAPTER I</b>	<b>I</b>
<b>Introduction and Motivation</b>	<b>I</b>
1.1. Problem .....	2
1.2 Contribution .....	3
<b>CHAPTER II</b>	<b>6</b>
<b>Literature Survey</b>	<b>6</b>
2.1 Related Work .....	6
2.2 Methodic Background .....	11
2.2.1 Data Structure as “Triples” .....	11
2.2.2 Serving Linked Data .....	13
<b>CHAPTER III</b>	<b>17</b>
<b>Data Management</b>	<b>17</b>
3.1 Structural Data Management .....	18
3.2 Content Based Data Management .....	20
3.2.1 Disease Dataset .....	21
3.2.2 Geographical Dataset .....	22
3.2.3 Statistical Data .....	23
<b>CHAPTER IV</b>	<b>29</b>
<b>System Overview</b>	<b>29</b>
4.1 Multi Layer Architecture .....	31
4.1.1 Data Layer .....	32

4.1.2 Service Layer.....	34
4.1.3 Presentation Layer .....	36
4.2 User-Access Methods .....	37
<b>CHAPTER V</b>	<b>40</b>
<b>Conclusions and Future Work</b>	<b>40</b>
5.1 Summary.....	40
5.2 Limitations and Lessons Learned .....	41
5.2.1 Limitations.....	41
5.2.2 Lessons Learned.....	42
5.3 Future Work .....	43
5.3.1 Utilizing OGC Services .....	43
5.3.2 Linking with Environmental Data .....	43
5.3.3 Enriching the Statistics .....	44
<b>Bibliography.....</b>	<b>45</b>



## Index of Figures

Figure 1: LOD Clod Diagram.....	7
Figure 2: Triples .....	12
Figure 3: Graph of 3 Statements .....	12
Figure 4: Database Schema .....	18
Figure 5: Logical Data Model .....	19
Figure 6: Conceptual Data Model.....	20
Figure 7: Workflow Diagram of Methodology.....	27
Figure 8: RDF-RDBMS Model.....	29
Figure 9: Multilayered Architecture .....	33
Figure 10: SPARQL Query Service Interface (4) .....	34
Figure 11: SOA Operations of a Linked SDI .....	35
Figure 12: Faceted browsing of a disease between different sources (1) .....	36
Figure 13: Representation of a spatial linked data with OpenLink Data Explorer (2) ....	37
Figure 14: SIMILE Exhibit data representation from Virtuoso .....	39

## Index of Tables

Table 1: Examples of location specific health information.....	3
Table 2: Domain Specific Ontologies.....	13
Table 3: Comparison of Triple Stores .....	14
Table 4: Example of Slices for Time Series .....	26

## Index of Listings

Listing 1: Dimensions .....	24
Listing 2: Measures .....	25
Listing 3: Attributes .....	25
Listing 4: Observation .....	25
Listing 5: Prefixes to define regions/cities dataset.....	30
Listing 6: RDF View Definition for regions dataset .....	30
Listing 7 : RDF Entity Mapping of regions dataset.....	30
Listing 8: Associations of regions dataset's entities with other entities .....	31

# CHAPTER I

---

## Introduction and Motivation

Health is an indispensable attribute of human life. Complex methodologies and multidisciplinary studies are needed to understand the relations between human health and environment. Environment plays an important role in health and human development within the context of genetic diversity to causation and variation in health - disease relation. Acute effects from exposure to environmental contaminants are well recognized and the relations are underlined by researchers. Exposure related specific environmental hazards with a health effect are exemplified as; benzene and leukemia or exposure to mixtures of drinking water disinfection with bladder cancer (Mather, et al., 2004). Pew Environmental Health Commission (2000) calls the lack of information linking environmental hazards and chronic disease the “environmental health gap”. For closing this gap, data from different disciplines has to be available thus a methodology can be developed for linking disparate datasets. Furthermore this methodology can be designated with interoperability at syntactic and semantic level with current technologies. Users and experts need appropriated information infrastructures, to exploit the potential of cross-disciplinary topics. This can be achieved by expanding the limited guidance for using available data from different disciplines such as health, environmental, socio-economic, statistical, and geospatial technologies and semantic web communities. Linked data, refers to “a set of best practices for publishing and connecting structured data on the web” accelerates the pace of discovery (Bizer, Heath, & Berners-Lee, 2009) with the main aim of integrating these resources in a transparent and communicable way so that they can be used to identify trends and relevant events.

In many research fields - from genetics and molecular biology to social sciences - data sharing is already ingrained in how researchers work (McCarthy, Abecasis,

& Cardon, 2008). The early example of using geographical information by John Snow to show the relation between water supply and cholera outbreaks in London, 1854 achieved by use of public data to link between contaminated water and disease (Johnson, 2006). Power of collaborative data is widely accepted by raise of crowd sourcing (Goodchild, 2007). OpenStreetMap<sup>1</sup> develops a freely accessible and editable digital map of the world by crowd sourcing and in the aftermath of the Haiti earthquake of 2010 hundreds of volunteers worked together to create a detailed map (Goodchild, 2011). Human genome project is also completed by global cooperation based on sharing data (HGP, 2003). In many countries, especially in USA<sup>2</sup>, UK<sup>3</sup>, and in some other EU Countries, formally sharing data on governmental or non-governmental platforms is becoming a common practice (Heath & Bizer, 2011). Noticeable importance and prevalence of public data, crowd-sourcing data and open data, lead us to underline the significance of open data integration with an ontology based approach for the context of the topic spatio-temporal health statistics.

## 1.1. Problem

Hans Rosling, health-turned-statistician-public-speaker talks about a rhetoric disease called DbHd: Database Hugging Disorder in his open data speech at World Bank on May 2010 (The World Bank, 2010). Formerly to this talk World Bank cleared itself from this disease by making World Bank data publicly available (April, 2010). Despite the common practices, culture of open data is not widely embraced yet. Much of the infrastructures, technical standards, and incentives that are needed to support data sharing are lacking, and these data can hold particular sensitivities (Walport & Brest, 2011). Therefore aside from privacy, security and reliability issues of open data, technical constraints are elaborated in the scope of this study, referencing the fact that well-established repositories and tools enable researchers to access and interrogate shared data resources, and build on one another's work (Cochrane & Galperin, 2010). Disparate data sources, unstructured data and several data format are available on the web, as data silos require a common structured hierarchy for a global database in order to communicate each other. This necessity brought the idea of designing a virtual documentation system for data sharing on the sky and

---

<sup>1</sup> OpenStreetMap <http://www.openstreetmap.org/>

<sup>2</sup> <http://www.data.gov>

<sup>3</sup> <http://data.gov.uk>

applying to the web of data.

Initially, lack of open data is stated as a problem, which is an impediment to linking and integrating data sources. Open data movement exists as a valuable opportunity to rectify this situation (Bizer C. , Heath, Ayers, & Yves, 2001). Second and a stimulating problem is to lack of semantic interoperability when inserting or retrieving the available data. Grounding the conceptual data into semantic fields within the context of health, geography and statistics requires explicit definitions and relations. Establishing an ontology that is able to communicate between the different topics (health, geography, statistics) and large domains of objects is complex and daunting. Moreover an infrastructure with defined functionalities between this ontological component and geographic information systems to achieve interoperability by exchanging and using the information is not well defined.

## 1.2 Contribution

Semantic Web is the next generation of WWW, provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries (W3C, 2001). Location-based information means information that is immediately relevant, which is the essence of the Semantic Web (Boulos K. M., 2003). The concept that place and location can influence health is a very old and familiar idea in medicine. As far back as the time of Hippocrates (c. 3rd century BC), physicians have observed that certain diseases seem to occur in some places and not others (Boulos K. , 2002). Geographic, economic, wealth and environmental profiles of the location affect the prevalence of the diseases. Some of the location related health information and data across other disciplines is listed in Table 1. Hence location matters for health; the capability of Geographical Information Systems (GIS) for visualizing and analyzing spatial data is used, with ontology based approach and a semantic flavor regarding to *Linked Data Principles*.

Table 1: Examples of location specific health information

Location Specific Health/Medical Information
- Local disease rates.
- Life expectancy according to social-economical profiles.
- Mortality rates at certain locations with certain profiles.

- Addresses of local health care facilities.
- Local weather, pollen and air quality alerts.
- Local health risks and hazards.
- Targeted health education.
- Travelers' health information.
- Local health news.
- Local drugs/drug trade names and prices
- Health sciences articles published as a case study for a specific location

*Source: Extended table from International Health Geographics 2003, 2*

Producing linked data requires a comprehensive groundwork apart from choosing the strong semantic relations within the context of topic but also structuring the data, storing it as RDF triples and serving to SPARQL endpoints. Unlikely to traditional relational databases and tables, linked data has a different approach to the web of data. In order to encourage data providers, research agencies, governmental organizations to use of this technology; pilot models, examples and tutorials are needed and yet being developed and published by academic institutions, governments, private organizations and individual entrepreneurs (Belleau, Tourigny, & Rigault, 2008). Particular case studies for a certain scope of topics are more favorable in order to draw a path for similar cases. This study takes as a case of mortality rates data observed in region of Valencia for different districts, classified by gender and cause of death as diseases. Statistical data is aggregated with different trends based on time and space and combinations of these two variables.

Prospected contribution of this study is to convey complex health related information to public health decision makers, data consumers and inform the development of future research and application opportunities by effectively utilizing an infrastructure. For this reason in the following chapters we display the process of data handling, ontology designation for geographical, statistical and health disciplines; server side inauguration and introducing client side access points, tools and methods for consuming data on web and different platforms. Data management and ontology designation between cross-domains is to accomplish best practices of linkage between health outcome and location based information (hazard, exposure, environmental etc.). Server side inauguration for

the data to be served via different channels and applications is to integrate, present, and analyze data for a wide range of users with semantic technologies and build a base for future developments.

## CHAPTER II

---

# Literature Survey

In this chapter, firstly related work mentioned within the context of accompanying topics; Linked Open Data and semantic web, health, statistics and geographical information. Related work section mentions about the similar works, covering the study topics entirely or partially according to the relevance. Methodic background gives a brief description about the tools, technological terms and descriptions, which are a part of the study. Ontologies related to the domain of interests to the case study are elaborated in methodic background, which leads up to selecting ontologies for describing relations of data in the following chapter.

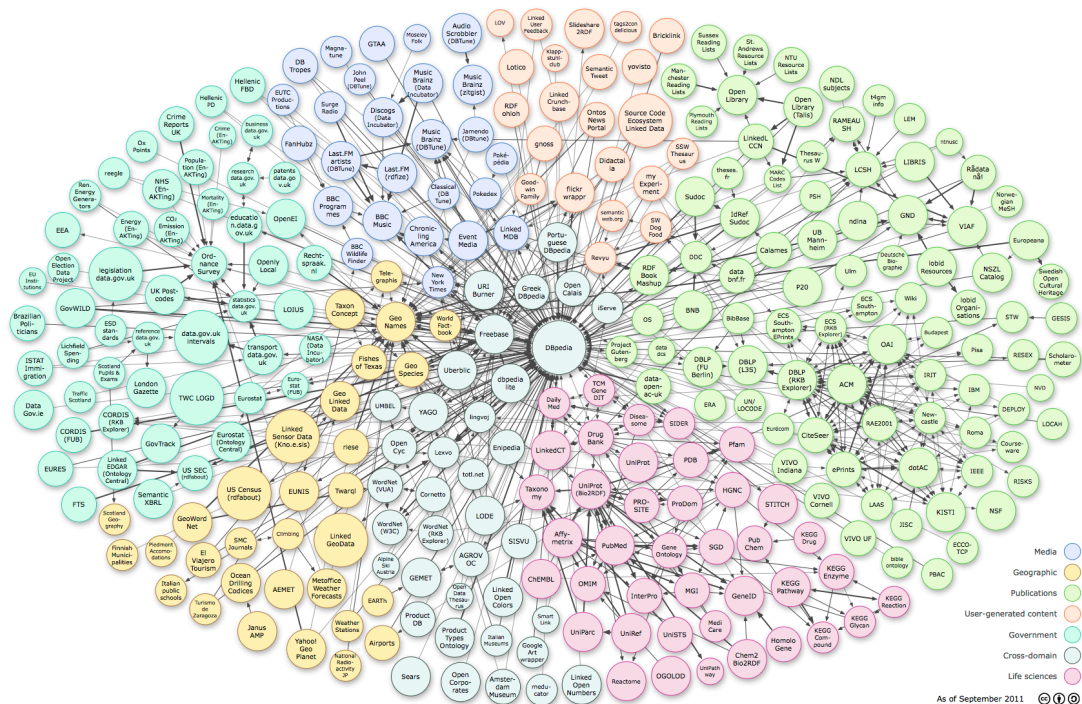
## 2.1 Related Work

The basic idea of Linked Data is to apply the general architecture of the World Wide Web (Jacobs & Walsh, 2004) and covering the need of publishing data on global space as stated in previous chapter's 1.1 Problem section. A set of best practices are outlined by *Linked Data Principles* introduced by (Berners-Lee, 2006) as follows:

1. Use Unified Resource Identifiers (URIs) as names for things.
2. Use Hypertext Transfer Protocol (HTTP) URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using standards Resource Description Framework, Sparql Protocol and RDF Query Language (RDF, SPARQL).
4. Include links to other URIs, so that they can discover more things.

Regarding to the principles stated above; as long as sources are identified uniquely they can be looked up with accepted common formats on a global scale. Followed by these principles global data space is created as Linked Open Data project where anyone can publish data to the web of data (Heath & Bizer, 2011). The web of data represented as a cloud diagram in Figure 1, which contains open-license datasets created by following linked data principles maintained by Comprehensive Knowledge Archive Network - CKAN<sup>4</sup> that can be explored through.

Figure 1: LOD Clod Diagram



Source: "Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. <http://lod-cloud.net/>"

Web of data classified into topical domains segregated by colors in the diagram above including geographic, government, life-sciences and cross-domain content. The data published as linked data that can be browsed through a web interface however still in fairly a raw format that makes reusability of those data possible. Datasets itself, which are related to the domains of this study, are interwoven with applications or demos created by re-use of such data hereinafter. Considering the reusability of the published linked data or potential to define the content of use case data specifies the methodology for selecting relevant examples.

<sup>4</sup> <http://www.ckan.net>



Semantic Web Health Care and Life Sciences Interest Group – **HCLSIG**<sup>5</sup> as an active community put a great effort for publishing several datasets as Linked Open Data in life sciences domain with topics of: Drugs, Clinical Trials, Diseases, Chemicals, Proteins, Genes, Side Effects, Drug interactions, Medicine Ingredients, Diagnoses etc. (HCLSIG, 2009). U.S. National Library of Medicine – **NLM**<sup>6</sup> under National Institutes of Health as the world's largest medical library provides information (i.e. Medical Subject Headings – **MeSH**<sup>7</sup>) and research services like National Center for Biotechnology Information – **NCBI**<sup>8</sup>. NCBI provides access to biomedical and health information through semantic resources like **PubMed**<sup>9</sup>, Online Mendelian Inheritance in Man – **OMIM**<sup>10</sup> and many others. Based on semantically structured and uniquely identified library sources a semantic data integration platform **LinkedLifeData**<sup>11</sup> for the biomedical domain is developed (Momtchev, 2009). DO - **Disease Ontology**<sup>12</sup> is another (opensource) project, semantically integrates these medical and disease vocabularies through extensive cross mapping between DO terms as well. **Diseasome**<sup>13</sup> is a network with a different range of semantic integration with disorders and disease genes linked by known disorder-gene associations for exploring all known phenotype and disease gene associations, indicating the common genetic origin of many diseases. Medical and health related information as thesaurus, libraries or similar sources enhance the meaning of existing disease data for the study case. Therefore mentioned sources above carry importance for the work and considered to be linked in further chapters.

Geospatial information needs to receive semantic specifications in order achieve interoperability (Kuhn W. , 2005). The potential of defining GIScience concepts and GIS data, explicitly with semantics lead to development of linked open data sets and semantic sources for GIS. **GeoNames**<sup>45</sup> is a frequently used data-hub that have geographical components published as linked open data. **GeoLinkedData**<sup>14</sup> is an open initiative of the Ontology Engineering Group (OEG) whose aim is to enrich the Web of Data with Spanish geospatial data. **LinkedGeoData**<sup>15</sup> is a wider scale effort to add spatial dimension to the web of

---

5 HCLSIG - <http://www.w3.org/wiki/HCLSIG>

6 NLM - <http://www.nlm.nih.gov/>

7 MeSH - <http://www.nlm.nih.gov/mesh/meshhome.html>

8 NCBI - <http://www.ncbi.nlm.nih.gov/>

9 PubMed - <http://www.ncbi.nlm.nih.gov/pubmed/>

10 OMIM - <http://www.ncbi.nlm.nih.gov/omim>

11 <http://linkedlifedata.com/>

12 <http://www.disease-ontology.org/>

13 <http://diseasome.eu/>

14 <http://geo.linkeddata.es/>

15 <http://linkedgeodata.org/>

data. It uses the information collected by the OpenStreetMap<sup>1</sup> project and makes it available as a knowledge base according to the Linked Data principles.

As well as geographical information, requires semantic definitions within its domain it can often connect information from varied topical domains. Medical Geography (Health Geographics) is one of these topics that bring interest to underline the related works in health and geography disciplines. The term "Geographic Information Systems" (GIS) has been added to MeSH<sup>7</sup> in 2003, a step reflecting the importance and growing use of GIS in health and healthcare research and practices (Boulos K. M., 2004). Evidently, the strong relation of health and geography as stated in 1<sup>st</sup> Chapter's Contribution section is demonstrated by following related works based on semantic web and linked data. **HealthMap**<sup>16</sup> (Brownstein, Freifeld, Reis, & Mandl, 2007) is an Internet-based system designed to collect and display information about new outbreaks according to geographic location, time, and infectious agent (Keller, 2009). **EpiSPIDER**<sup>17</sup> - **S**emantic **P**rocessing and **I**ntegration of **D**istributed **E**lectronic **R**esources for Epidemics (and disasters). Despite similarities between HealthMap and EpiSPIDER they monitor different data types and distribute information through different formats and ways. HealthMap aggregates data by source, disease and geographic location then overlays on an interactive map for user-friendly access to reports. It is designated for human consumption while EpiSPIDER redistribute aggregated data in a structured way for consumption of services, which can process semantically.

There exist a long list of semantic web applications that are deployed by consuming linked data, grounded with geographic information, directly or indirectly related to health. Governmental and domain specific applications bring together the interdisciplinary topics. Stirring up the "delineation of health-related data with spatio-temporal aspects and representing with statistical definitions on semantic web" summarizes the scope of the study. Related applications matches partially or completely to this summary is as follows:

- **Meteorological Sensor Data** represents the state of the atmosphere (humidity, pressure, temperature, wind, etc.) in Spain by using GeoLinkedData<sup>14</sup> model with the latest semantic web technologies (AEMET, 2010).

---

<sup>16</sup> <http://healthmap.org/>

<sup>17</sup> <http://www.epispider.net/>

- **Deprivation areas in England** is represented by a map application showing the lower layer super output areas in England, classified by deprivation indicators (housing, crime, health, environment, education etc.) developed by Open Data Communities (IMD, 2010).
- **Stats2RDF** project is carried by AKSW<sup>18</sup> – Agile Knowledge Engineering and Semantic Web research group. The whole topic of the project stated as “Representing multi-dimensional statistical data as RDF using RDF Data Cube Vocabulary”. World Health Organization – WHO’s Global Health Observatory<sup>19</sup> dataset is used as an initial use case (Zaveri A. J., 2011).
- **EnAKTing** is an EPSRC – Engineering and Physical Sciences Research Council<sup>20</sup> funded Advanced Knowledge Technologies Interdisciplinary Research Collaboration project (AKT IRC) one of the pioneers as Tim Berners-Lee<sup>21</sup> has a significant success and well established sources and services (EnAKTing, 2009). **EnAKTing PSI mortality dataset**<sup>22</sup> shows mortality statistics per region in England for the year 2008/09 represented through different navigation services. **Openspace Map Visualization**<sup>23</sup> shows possible applications of such services including mortality statistics. Geographical reasoning is a service to discover geographical resources in the Web of Data querying containment relations provided by **PSI Geographical Service**<sup>24</sup>. Containment API provides data, in addition to the local content, from several sources like DBPedia, GeoNames, Ordnance Survey, Open CYC etc.
- **GeoSpecies** knowledge base ties together disparate data about species like animals, plants, protozoans, viruses and such connecting by their location, soil attributes, climatological attributes, ecological zone types. As some species are pathogenic and cause diseases GeoSpecies is a promising knowledge base related to public health studies.
- Previously it’s mentioned in Introduction “governmental bodies and

---

18 AKSW - <http://aksw.org/>

19 GH0 - <http://apps.who.int/ghodata/>

20 EPSRC Grant - <http://gow.epsrc.ac.uk/ViewGrant.aspx?GrantRef=EP/G008493/1>

21 <http://www.w3.org/People/Berners-Lee/>

22 <http://mortality.psi.enakting.org/>

23 <http://map.psi.enakting.org/>

24 <http://geoservice.psi.enakting.org/>

public-sector organizations produce a wealth of data” (Heath & Bizer, 2011). **Data.gov**<sup>25</sup> here is exemplified with its mash-up map applications with linked data. **Clean Air Status Trends-Ozone**<sup>25</sup> shows a map of the U.S. depicting the Clean Air Status and Trends Network (CASTNET) Ozone monitoring stations and their average Ozone level. Clicking a station shows more detailed information and a link to a graphical representation of the time-series of measurements. **Trends in Smoking Prevalence, Tobacco Policy Coverage and Tobacco Prices (1991-2007)**<sup>26</sup> looks at how smoking rates, population, cigarette taxes, and other related variables relate to one another, by state in USA.

## 2.2 Methodic Background

In this section of the second chapter, literature and technologies are introduced regarding to linked data concept. First part gives a brief statement of the main points for constitution of triples and lists the domain related namespaces used for representing use case data. RDF creation is mentioned theoretically on an introductory level and to intricate pattern with use case data left for the next chapter. Second part compares triple stores according their performance and flexibility to integrate with third party applications. It is also briefly discussed the level of support for different types of data inputs from relational databases and web sources.

### 2.2.1 Data Structure as “Triples”

Linking Open Data project is a visible example of the integration of diverse data that covers also significant portions of health sciences; genes, proteins, drugs and clinical trials as well as statistical and geographical data (Bizer, Heath, & Berners-Lee, 2009). In order to be able to communicate with other data sources, use case data is supposed to be prepared by following the *Linked Data Principles* that are stated in the previous section at page 6. URIs - Unified Resource Identifiers are used for identifying the granular data. This approach gives more generic means to identify existing entities, attribute names, keys and relations. Regarding to the second principle, URIs are required to be constructed with `http:// scheme` which can be dereferenced on web by `http`

---

<sup>25</sup> <http://data.gov.tw.rpi.edu/demo/exhibit/demo-8-castnet.php>

<sup>26</sup> [http://logd.tw.rpi.edu/demo/trends\\_in\\_smoking\\_prevalence\\_tobacco\\_policy\\_coverage\\_and\\_tobacco\\_prices](http://logd.tw.rpi.edu/demo/trends_in_smoking_prevalence_tobacco_policy_coverage_and_tobacco_prices)

protocol. HTTP URI carries importance for establishing the single data model as RDF (Resource Description Framework) for publishing structured data on the web. Whilst HTML as a dominant document format linking documents on the web, RDF stands for linking data on web. At the simplest level, the Resource Description Framework is an XML-based language to describe resources (Daconta, Obrst, & Smith, 2003). It allows to formulate statements about resources, each statement consisting of subject, predicate, object. Description of the data in these statements usually refers to form of triples as subject – predicate – object. Subject is a URI, object can be a URI as well or a literal and predicate is usually a URI defining the relations between subject and object (Figure 2). Whilst an object is a URI can be a subject of another object linked with predicates (Figure 3). Predication is to say something about the subject and a good practice for choosing predicates is considered as, reusing existing ontologies and vocabularies like RDF/RDFS (Resource Description Framework/Schema), OWL (Ontology Web Language), SIOC (Semantically Interlinked Online Communities), FOAF (Friend of a Friend), Dublin Core etc. in order to define interoperable relations (Bizer, Heath, & Berners-Lee, 2009).

Figure 2: Triples

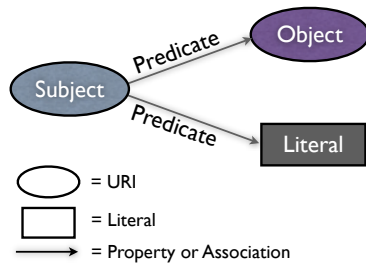
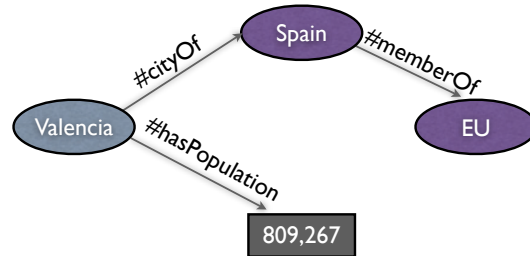


Figure 3: Graph of 3 Statements



Gruber has given the widely cited definition of ontology: “ontology is an explicit specification of conceptualization” (Gruber, 1993). Ontologies, schemas and vocabularies, all mean roughly the same thing, which are defining RDF information about RDF information (Tauberer, 2005). As to the context of study topic that clarified in motivation section that covers spatio-temporal, health statistics are the main disciplines defining the use case data. In order to convert the data into RDF triples, existing ontologies and vocabularies are elaborated, grouped and listed in Table 2.

Table 2: Domain Specific Ontologies

Domain	Ontology	Predicate
Spatial	prefix gn: < <a href="http://www.geonames.org/ontology#">http://www.geonames.org/ontology#</a> >	Interlinking with <owl:sameAs>
	prefix geo: < <a href="http://www.w3.org/2003/01/geo/wgs84_pos#">http://www.w3.org/2003/01/geo/wgs84_pos#</a> >	geo:lat, geo:long
	prefix dcterms: < <a href="http://purl.org/dc/terms/">http://purl.org/dc/terms/</a> >	dcterms:location
	prefix tisc: < <a href="http://observedchange.com/tisc/ns/#areasize">http://observedchange.com/tisc/ns/#areasize</a> >	tisc:areasize
Time	prefix time: < <a href="http://www.w3.org/2006/time#">http://www.w3.org/2006/time#</a> >	time:year
Health	prefix MeSH: < <a href="http://www.nlm.nih.gov/cgi/mesh/2012/">http://www.nlm.nih.gov/cgi/mesh/2012/</a> >	Interlinking with <owl:sameAs>
	prefix Diseases: < <a href="http://www4.wiwiwiss.fu-berlin.de/diseases/resource/diseases/">http://www4.wiwiwiss.fu-berlin.de/diseases/resource/diseases/</a> >	
	prefix dbpedia :< <a href="http://dbpedia.org/resource/">http://dbpedia.org/resource/</a> >	
Statistical	prefix qb: < <a href="http://purl.org/linked-data/cube#">http://purl.org/linked-data/cube#</a> >	qb:dimension,
	prefix scovo: < <a href="http://purl.org/NET/scovo#">http://purl.org/NET/scovo#</a> >	qb:slice, qb:item

The Unique Name Assumption (UNA) is a concept from ontology languages and description logics. In logics with the unique name assumption, different names always refer to different entities in the world (Russell & Norvig, 2003). Instead of making this assumption Ontology Web Language (OWL) constructs an explicit predicate to express that two names express the same things in a mutually compatible way. For this purpose `owl:sameAs` is the OWL predicate that, asserts two URIs refer to same entity. As stated in the table, existing ontologies that contain the same thing, which is referring to the same entity in the use case data are interlinked by `owl:sameAs`. This linkage maintains to reference for the same data object from different information providers.

### 2.2.2 Serving Linked Data

Triple or RDF graphs that are small can be efficiently handled and managed in computers or servers main memory; larger RDF graphs render the deployment of persistent storage systems indispensable. Data volume – size of data to be served and data dynamism – frequency rate of updating data are main considerations for establishing a triple store. Triple stores are purpose-built databases for the storage and retrieval of any kind of data expressed in RDF (Haslhofer, Momeni, Schandl, & Zander, 2011). Triple stores are defined as; “RDF stores as systems that allow ingestion of serialized RDF data and the

retrieval of these data later on” in Europeana<sup>27</sup>’s RDF Store Report. Before starting this study, investigation on RDF stores has been a compulsory step. Due to the peculiarities of RDF stores compared to relational database models ease of use in terms of installation, development and administration carries importance as well as supporting interoperability. Related to the scale and size of the data query performance and costs carry importance for linked data entrepreneurs. For instance, according to the triple store evaluation analysis report of Revelytix<sup>28</sup> two best stores in terms of efficiency and cost-effectiveness are Ontotext’s BigOWLIM<sup>29</sup> and Openlink Software’s **Virtuoso Open Source Edition**<sup>30</sup> (Revelytix, Inc. , 2010).

Table 3: Comparison of Triple Stores

	Area	Notes	Virtuoso	OWLIM	Oracle	Mulgara	Parliament	AllegroGraph
General	Usability	Overall perception of usability						
	Support	Overall perception of support						
	Licensing/Cost	Assessment of licensing/cost						
Functional	System	System architecture, etc						
	Enterprise	Enterprise capabilities						
	Data Import	Ability to import data						
	API	APIs available						
	Querying	Query functionality						
	Inferencing	Inferencing support						
	Interoperability	Jena/Sesame interop						
Performance	Operational	Monitoring/canceling						
	Speed	Single-user, SP2						
		Single-user, BSBM						
	Throughput	Multi-user, SP2						
	Correctness	SP2 queries/data						
		BSBM queries/data						

LEGEND: Green is 'very good', red is 'not so good', and yellow is in-between. Gray is unknown.

Source: Revelytix Triple Store Evaluation Analysis Report, p6.

General classifications of RDF stores are entitled as; native stores, DBMS backed stores and hybrid stores. Native stores implement a complete DB engine on their own and doesn’t reuse the retrieval functionalities and storage models of other DBMS while DBMS backed stores use these functionalities and storage models (triple tables in DB, ontology-aware models). Hybrid stores support both architectural styles (native and DBMS-backed). **OWLIM**<sup>29</sup> offers a family of native RDF store solutions preferable according to data volume and free or commercial licenses. RDF stores are also solution mechanisms for controlling frequent changes in the data. If the data is stored in a relational database the replication of the RDF store required to be handled through an RDB-to-RDF wrapper. RDF wrappers are lightweight software components laid on top of existing data source. **D2R**<sup>31</sup> Server provides nice tools for this RDF-wrapping

27 <http://www.europeana.eu/>

28 [www.revelytix.com](http://www.revelytix.com)

29 OWLIM - <http://www.ontotext.com/owlim>

30 Virtuoso - <http://virtuoso.openlinksw.com/dataspace/dav/wiki/Main/>

31 D2R - <http://www4.wiwiwiss.fu-berlin.de/bizer/d2r-server/>



mechanism by using **D2RQ**<sup>32</sup> platform and mapping language. It threats non-RDF; relational databases as virtual, read-only RDF graphs and offers a variety of different RDF-based access mechanisms to the content of huge, non-RDF databases without having to replicate the database into RDF (Bizer, Cyganiak, Garbers, Maresch, & Becker, 2009). **SquirrelRDF**<sup>33</sup> and **METAmorphoses**<sup>34</sup> are some other tools that initiate relational data to RDF mapping. An example of a hybrid store is Virtuoso for a range of data models including relational data, XML, CSV, RDF and free text documents. Through it unified storage it can be also seen as a mapping solution between RDF and other data formats, therefore it can serve as an integration point for data from different, heterogeneous sources (Erling & Mikhailov, 2007). Considering the data integration from multiple heterogeneous sources and capability of handling data dynamism Virtuoso seemed to be the most promising triple store to be established within scope of this study. Virtuoso server can be also accessed through a number of client APIs (native JDBC<sup>35</sup> interface, SPARQL endpoint interface, Jena<sup>36</sup>-based interface) that is an important feature to develop applications later on. Virtuoso provides a built-in Linked Data interface and support for SPARQL query language where the administrator can configure the content of the store to be published. Ideally a triple store would provide a Linked Data interface whilst not **Pubby**<sup>37</sup> is a widely used interface that can serve on top of triple store's SPARQL endpoint. Virtuoso as hybrid store has it's own relational database and SQL query language, which supports geometry data type ands and spatial indexes. Through SQL built-in functions geometry data types can be queried by using SPARQL. A SPARQL function is provided to convert a pair of latitude and longitude property values into a point geometry. A special literal data type, `virtrdf:Geometry`, is also provided for indexing point literals. Support for testing intersection and containment relationships is provided via property functions (Battle & Kolas, 2011). To define and publish complex geometry types as linked data can be done in different ways. If the data is being extracted from a relational database SQL/MM<sup>38</sup> geometry data type can be defined in RDF and queried by applying SQL MM geometry predicates in SPARQL with Virtuoso. This special literal data type mentioned above `virtrdf:Geometry` handles this

32 **D2RQ** - <http://www4.wiwiwiss.fu-berlin.de/bizer/d2rq/>

33 **SquirrelRDF** - <http://jena.sourceforge.net/SquirrelRDF/>

34 **METAmorphoses** - <http://metamorphoses.sourceforge.net/>

35 **Java Database Connectivity** for Virtuoso: <http://docs.openlinksw.com/virtuoso/VirtuosoDriverJDBC.html>

36 **Jena** is a Java framework for building Semantic Web applications. Jena provides a collection of tools and Java libraries to help you to develop semantic web apps, tools and servers. Virtuoso **Jena** Provider: <http://openlinksw.com/JenaProvider>

37 <http://www4.wiwiwiss.fu-berlin.de/pubby/>

38 ISO/IEC 13249 **SQL/MM** is the effort to standardize extensions for multi-media and application-specific packages in SQL (Stolze, 2003).



process. Another solution for relational data is offered by (Valle, Qasim, & Celino, 2010) with a prototypical system (namely G2R) and a declarative mapping language as an extension of D2RQ<sup>32</sup> mapping language with support of geometry data type in relational databases as well as a prototypical system (namely G2R) that allows SPARQL query involving spatial computation to be executed in a mixed environment.

An upcoming OGC – Open Geospatial Consortium standard **GeoSPARQL**<sup>39</sup> attempts to unify data access for the geospatial semantic web. GeoSPARQL defines a vocabulary for representing geospatial data in RDF, and defines an extension to the SPARQL query language for processing geospatial data. Another apprising initiative is **map4rdf**<sup>40</sup> that serves as a mapping and faceted browsing tool for exploring and visualizing RDF datasets enhanced with geometrical information. Previously mentioned GeoLinkedData<sup>14</sup> dataset is published with using map4rdf technology. Rising geospatial semantic standards and tools for spatial RDF data are available among a range of possibilities though not sufficient yet to be challenging to **SDIs**, Spatial Data Infrastructures<sup>41</sup> however both are versatile to be complementary. Bootstrapping Linked Open Data with SDI is an outset for OGC and LOD communities. Apprising studies and efforts are engaged in this process (Battle & Kolas, 2011) (Pehle, Bootstrapping the Web of Linked Locations, 2010) (Pehle, The Role of Government SDI in Linked Data, 2011).

---

<sup>39</sup>GeoSPARQL - <http://www.opengeospatial.org/projects/groups/geosparqlswg>

<sup>40</sup> map4rdf - <http://oegdev.dia.fi.upm.es/projects/map4rdf/>

<sup>41</sup> An **SDI** is a coordinated series of agreements on technology standards, institutional arrangements, and policies that enable the discovery and use of geospatial information by users and for purposes other than those it was created for (Kuhn W. , 2005).

## CHAPTER III

---

# Data Management

Data management for establishing a semantic system is the backbone of the study. Efficient management of RDF data is an important factor in realizing the semantic web vision (Abadi, Marcus, Madden, & Hollenbach, 2007). A substantial block in semantic vision is the necessity of modeling web data conceptually. Today the vast bulk of data held by companies resides in relational databases and, as a result, data that ultimately reaches the web is inherently heterogeneous at both the data schema and Database Management System-DBMS engine levels (Virtuoso, 2010). For that reason a key infrastructural requirement of the semantic web vision is treating non-RDF relational data as virtual RDF graphs and facilitating the generation of RDF views of relational data. Thus preparation of semantic links based on relational data is eased.

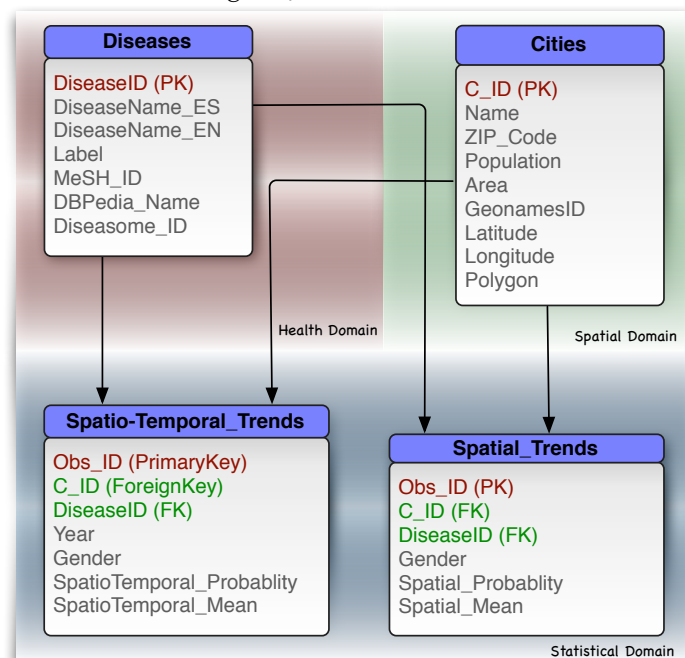
This chapter describes preparing data for adaptation to semantic web in order to achieve semantic interoperability. As stated earlier common format for linked data is triples. As a result of this data management process RDF links as triples are produced. When RDF links are being created, semantics of the data objects are added by previously introduced domain specific ontologies to describe each data objects.

Initially relational data model is introduced and progress of RDF mapping is explained. In the second section conceptual data management is explained by elaborating the data in three different conceptual spaces (spatial, statistical, health) according to the content of the data. Semantics between different domains of data are explained by figures and listings. Final structure of the data is available to be used in an SDI web services, linked data browsers and applications. Creating, managing and handling the data in multi-structured, communicable and interoperable way will leverage the future research opportunities with 3<sup>rd</sup> party datasets.

## 3.1 Structural Data Management

Use case data is gathered from CSISP - Centro Superior Investigación en Salud Pública; Valencia Public Health Research Center. Initial data is received in raw R-data<sup>42</sup> format as several files. The data was an output of statistical calculations exported in R-data format. It didn't contain anything than numbers encoded according to placeIDs and grouped into 20 years of time span. It was organized into those several files according to some key attributes like gender and diseases. What is emphasized here is that the data was syntactically structured but not semantically. Such kind of data doesn't allow to be consumed by others due to lack of metadata and is not suitable for interlinking with other datasets. The data is clear just for the ones who produce and analyze it however limited for cross-domain analyses and usage. The only way to consume this data is to visualize on graphs, charts and maps dynamically without the ability to be upheld. First step of data management is to carry the data up to advanced levels and show the bidirectional functioning between those. Due to pervasiveness of relational databases a logical schema is drawn and data transferred to a relational database. Thus, RDF scheme can be created by using this relational data and keep updated. A simple database schema is drawn in the following figure.

Figure 4: Database Schema



<sup>42</sup> Rdata is a lazy load database data that provides an object related information dynamically when it's actually needed by instantiating the related objects and its properties when it's requested.

Tables in the database are created according to the domains; spatial, health and statistical where spatial domain table is populated by cities of Valencian Community cities. Mortal diseases observed in the cities of Valencian community and their attributes populate health domain table. Attribute fields storing the ID's of the disease from vocabularies like MeSH<sup>7</sup>, Diseasesome<sup>13</sup> are futile in the relational database though when they are converted to URI's of the same disease in the RDF data, they become convenient and bring more information that can't be stored in the local relational database. Statistical domain tables are based on spatio-temporal trends with 20 years of time span and just spatial trends as an average of these years. Some sample fields of the tables stored in the database can be seen below in Figure 5: Logical Data Model.

*Figure 5: Logical Data Model*

CITIES TABLE								
C_ID	Name	ZIP_Code	Population	Area	GeonamesID	Latitude	Longitude	Polygon
77	Benicassim	12028	9037	36039979.8	6356985	40.0577061	0.04280985	"0.08,40.08 ...

DISEASES TABLE						
DiseaseID	Label	DiseaseName_ES	DiseaseName_EN	MeshID	DBPedia_Name	DiseasomeID
16	Colon	Tumor maligno de colon	Colonic Neoplasms	D003110	Colorectal_Cancer	1914

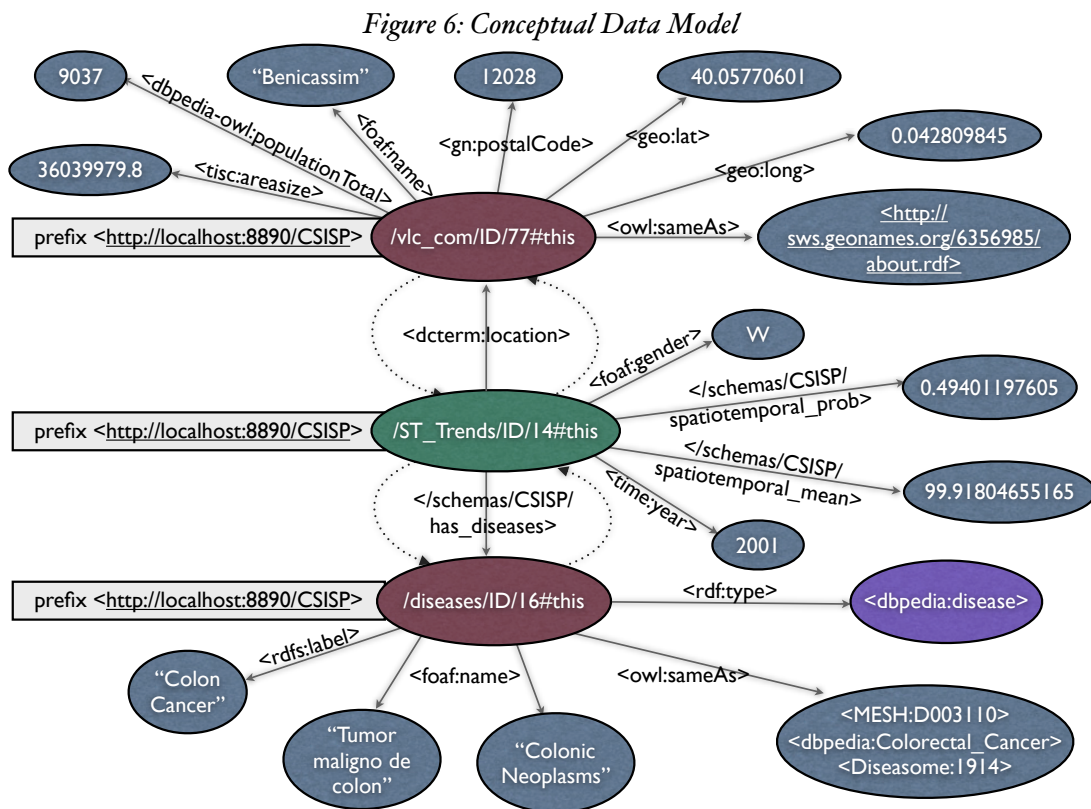
SPATIO-TEMPORAL TRENDS TABLE							
Obs_ID	C_ID	DiseaseID	Year	Gender	SpatioTemporal_Probability	SpatioTemporal_Mean	
14	77	16	2001	W	0.49401197605	99.91804655165	

The problem domain is expressed in logical schema within relation to entities through primary key and foreign keys as the observations in spatio-temporal trends table are for certain cities, for specific diseases. However semantic description of entities in problem space is lacking when data is subject to store in relational databases. This relational schema can be rendered to RDF by assigning primary keys as unique identifiers, which will serve as subject of RDF triples whereas they are foreign keys of a table then will serve as object of the triple. Subjects and objects are linked through meaningful URIs called predicates given in Table 2: Domain Specific Ontologies, raise the semantic description. Each entity in the tables is identified by URI's instead of their names or primary keys. As primary keys are non-repeating and unique, instead of using names in `http://` URIs primary keys are used as IDs. Therefore any character encoding problem is also prohibited when URIs are looked up. Initially, namespace where the URIs will be located on the server is described in the domain `/CSISP`. Finally minting URIs under the namespace CSISP is done and classified according to the range of table names; diseases, cities and spatial

trends. Sample data in Figure 5: Logical Data Model can be observed as URIs in Figure 6: Conceptual Data Model.

## 3.2 Content Based Data Management

Relationships between two entities are not explicit in the model at Figure 1. Foreign keys relating two tables don't precisely express the nature of relationship. Moreover this schema belongs to specifics of a particular vendor's RDBMS and in heterogeneous database environment must handle different query language dialects and perpetual problems for integration external datasets. This paradigm shift from logical models to conceptual models aimed to provide better and



explicit semantics from "search" to "esoteric precision find" and easier heterogeneous data integration. As semantic technologies provide building blocks for conceptual models and vocabularies define the concepts and their relationships in a domain of interest, in the subsections, data is elaborated according to its domain of interests.

### 3.2.1 Disease Dataset

Aside from representing common semantics also, vocabularies are a key part of Linked Data principles as they provide means to overcome semantic interoperability problems (Bishr, 1998). The necessity of vocabularies in domain of health can be exemplified by referring to the interpretation of a disease name in medicine terminology and different languages. The term cancer in medicine terminology refers to carcinoma, malignant tumor (PubMed, 2010), neoplasms (MeSH, 2012) which is the disease caused by an uncontrolled division of abnormal cells in a part of body (NOAD, 2005). The same term is used with different meanings in astronomy and astrology<sup>43</sup>. Overcoming such ambiguities can be achieved by taxonomy of concepts and publishing them as vocabularies for specific domains thus they can be reused through unique identifiers, hence semantic interoperability is provided.

Mortal diseases observed in the dataset exist as linkable data objects through URIs on the web of data in MeSH Vocabulary, DBPedia and Diseasesome network among many others. Linking the data to these existing vocabularies with the predicate `owl:sameAs` enriches the local data. By the power of this predicate two data which doesn't know each other with a link reference will implicitly known by the rest of the data linked to the subject. Heritable disorders of the disease, related diseases and medicine, causes and treatments, associated genes, succinctly etiology of the disease are some of the data in which can be extracted from these three vocabularies linked to the use case data. Attribute values; ID's and unique names that are stored in the fields of the diseases table at Figure 5: Logical Data Model for health related vocabularies, are appended to their prefixes, which is listed under the health domain at Table 2: Domain Specific Ontologies.

In the domain of health, apart from interlinking the URIs created with the use case data containing 27 mortal diseases observed no semantic relations are made by using ontological descriptions like symptoms of the disease, related drugs, disorders and such. Such datasets and relations can be created by domain experts and health scientists. Use case data contains only the name of the diseases, for that reason a way for enriching the local context is chosen by linking to rich domain vocabularies.

---

<sup>43</sup> *Astronomy*: a constellation (the Crab), said to represent a crab crushed under the foot of Hercules. *Astrology*: the fourth sign of the zodiac, which the sun enters at the northern summer solstice. (New Oxford American Dictionary, 2005).

### 3.2.2 Geographical Dataset

Spatial data is provided as shape file, which is a common format of GIS data (ESRI, 1998). Dataset covers the Valencia, an autonomous community<sup>44</sup> defined as a first level administrative division with its 541 cities / municipalities which are third level administrative divisions. Place name (toponym) information differ in the quality of their associated data, such as the feature types (e.g. city vs. municipality, community vs. state) and the spatial footprints (i.e. reference points, coordinates), there is need for a form of conflation (Hastings, 2008). For this reason multiple sources are merged so that the different aspects of each source can be exploited. GeoNames<sup>45</sup> is chosen to start with enriching spatial data. GeoNames represents each feature in the dataset as web resource through a stable URI. Data extraction from GeoNames dataset is limited to Valencian Community with its third level administrative divisions, cities into local disk as text files. String matching algorithms are used for corresponding feature IDs with place names in the local dataset and insertion of the IDs to the “*GeonamesID*” field are completed as in Figure 1, cities table. Soundex (Knuth, 1998) is used which is a phonetic algorithm for matching similar sounding names by converting them to the same code and resulted as almost fully match for 540 cities of 541. RDF links connected with the same predicate `owl:sameAs` used for linking disease names is used for interlinking place names. URIs are normalized by appending “*GeonamesID*” field to the namespace of GeoNames (<http://sws.geonames.org/>) in order to be matchable with the URIs represented for GeoNames dataset (Figure 2). Consequently, this record is augmented with population, postal code, area size, polygon coordinates and more global definition of place type from a matching GeoNames entry. GeoNames is particularly rich in town/city scale features as well as natural geographic features such as rivers, lakes, mountains, coasts and valleys (Smart, Jones, & Twaroch, 2010). Possibility of relating local data with such natural geographic features carries importance for relating health outcomes and exposure / hazard data aggregated over a geographic area for further objectives of the study. In addition to interlinking and enriching local data with GeoNames dataset, geospatial information in the local data needs to receive semantic specifications for semantic interoperability.

---

44 An autonomous community (Spanish: ‘comunidad autónoma’) is the first-level political division of the Kingdom of Spain established in accordance with the current Spanish constitution (1978). (Wikipedia, 2012).

45 GeoNames is a geographical database that is available for download free of charge under a creative commons attribution license. <http://www.geonames.org>

Geospatial semantics is about understanding GIS contents, and capturing this understanding in formal theories (Kuhn W. , 2005). Contents of geographical dataset in the case of this study are basic elements like area of the place, number of inhabitants on this place and geographical footprints of the place for instance coordinates in latitude/longitude and polygon coordinates. Referencing to another definition of ontology as a “logical theory accounting for the intended meaning of a formal vocabulary” (Guarino, 1998), ontologies are the tools for capturing the meanings of GIS contents. Codifying the relations of the GIS contents with its main geographical element in this case cities of the Valencian Community by using formal vocabularies are depicted as a graph schema in Figure 6: Conceptual Data Model. For that matter, ontologies listed earlier in Table 2: Domain Specific Ontologies for spatial domain is used for describing the contents of the dataset. Area size, postcode, population, coordinates of the cities are described by these formal vocabularies. Semantic descriptions for polygonal boundaries can be provided with an amalgamated approach by introducing SQL/MM<sup>38</sup> data type during creation of RDF and applying SQL/MM geometry predicates for retrieving data in SPARQL.

### 3.2.3 Statistical Data

Statistical data is of paramount importance in this data integration study. Semantics of the statistical data or in other means RDF representation is fairly different than usual methods. Understanding the statistical terms and representing them with formal vocabularies with a semantic flavor to the geographical health data is the objective in this section.

Third table in the Figure 5: Logical Data Model, for spatio-temporal trends include statistical data aggregated by inference models i.e. Bayesian spatio-temporal models. A single statistical data item in the dataset is treated with in a multi-dimensional way by using statistical vocabularies; *cube* and *scovo* as listed in Figure 6: Conceptual Data Model in domain of statistics. Despite the fact there has been studies, frameworks and tools based on scovo vocabulary (Hausenblas, et al. 2009), (Zaveri, et al. 2010) they announced on their homepage<sup>46</sup> as it is deprecated and directs to use RDF Cube Vocabulary (Cyganiak, Reynolds, & Tennison, 2010).

The cube model consists of three main components; dimensions, measures

---

<sup>46</sup> The Statistical Core Vocabulary (SCOVO) - <http://vocab.deri.ie/scovo>



and attributes. Dimensions identify statistical observations in terms of concepts like gender, region, time and such. Measures represent the phenomena observed and attributes qualify and interpret the observed values in terms of the units of measures or scaling factors. Elaborating these components based on the use-case sample of statistical data set occurs as follows:

- Dimensions; Cities of Valencia region, 20 year time span and gender {F, M}
- Measure: Mortality rates
- Attributes: Probability distribution and mean calculation methods or unit measures of spatio-temporal trends. (Optional)

Creating data structure definitions for spatio-temporal trends table shown at Figure 5: Logical Data Model is sorted in the following listings. Before listing the components additional concepts to introduce are based on SDMX - Statistical Data and Metadata Exchange<sup>47</sup>. This standard provides content oriented guidelines, which define a set of common statistical concepts and associated code lists that are intended to be reusable across data sets as cube vocabulary uses in the example. Listing 1 is an example of how to define dimensions as RDF links. Abbreviation used for cube vocabulary is "qb." New predicates defined in domain of CSISP are represented within the range of previously relations depicted in Figure 6: Conceptual Data Model;

dcterms:location, foaf:gender, owl-time:year

*Listing 1: Dimensions*

```
<CSISP:vlc_com>
  rdf:Property qb:DimensionProperty;
  rdfs:subPropertyOf sdmx-dimension:RefArea;
  qb:concept sdmx-concept:RefArea;
  rdfs:range dcterms:Location.
<CSISP:Gender>
  rdf:Property qb:DimensionProperty;
  rdfs:subPropertyOf sdmx-dimension:sex;
  rdfs:range foaf:gender.
<CSISP:time>
  rdf:Property qb:DimensionProperty;
  rdfs:subPropertyOf sdmx-dimension:refTime;
  qb:concept sdmx-concept:RefTime;
  rdfs:range owl-time:year.
```

---

<sup>47</sup> <http://sdmx.org>

Next listing describes component; measure is to give the value of each individual observation attached by different observed values for different calculation methods (probability distribution and mean) for the represented phenomena; Mortality rates.

*Listing 2: Measures*

```
<CSISP:MortalityRate>
  rdf:Property qb:MeasureProperty;
  rdfs:subPropertyOf sdmx-measure:obsValue;
  rdfs:range xsd:decimal;
  qb:measure CSISP-measure:stProbablity;
  qb:measure CSISP-measure:stMean.
```

Optionally attributes, which clarify the unit measures or statistical methods for measures, are given as an example in Listing 3. These components are used to redefine dataset with stronger semantic relations which are likely consistent with other statistical domains. Finally in Listing 4, sample data is shown as a qb:observation by its dimensions and measures.

*Listing 3: Attributes*

```
<CSISP-measure:stProbablity>
  rdf:Property qb:AttributeProperty;
  qb:attribute dbpedia:Probability_distribution.
<CSISP-measure:stMean>
  rdf:Property qb:AttributeProperty;
  qb:attribute dbpedia:Mean.
```

*Listing 4: Observation*

```
</ST_Trends/ID/14/this#>
  rdf:Property qb:Observation;
  qb:Dataset CSISP:ST_Trends;
  CSISP:vlc_com </vlc_com/ID/77/this#>;
  CSISP:Gender sdmx-code:sex-W;
  CSISP:Time 2001;
  CSISP-measure:stProbablity 0.49401197605;
  CSISP-measure:stMean 99.91804655165.
```

Cube vocabulary supports also creating slices by grouping observations with fixed dimensions to reduce the verbosity of the dataset and guide consuming applications (Cyganiak, Reynolds, & Tennison, 2010). An example of creating slices would be by creating time series from ST\_Trends table. Another

dimension of the slice can be chosen as regions. This way of designation of the slice allows to note a change in measurement process which affects a particular time or region<sup>48</sup>. To narrow down the slice one specific diseases chosen that previously used in examples at Figure 5: Logical Data Model and Figure 6: Conceptual Data Model “Colon Cancer” and listed for 3 sample cities as follows at Table 4: Example of Slices for Time Series. We can see the 3 Dimensions; region, time, gender which are listed at Table 4 are shown bold. The data is sliced through time series and the dimensions below time slices are sections. Each observation represents the spatio-temporal probability values of mortality rates for colon cancer in listed regions. Adding sections below gender dimension and showing different statistical methods for aggregated numbers can enhance statistical observation values.

*Table 4: Example of Slices for Time Series*

<i>2001-2006 stProbability Values of Colon Cancer</i>	<b>2001</b>		<b>2002</b>		<b>2003</b>		<b>2004</b>		<b>2005</b>		<b>2006</b>	
	<b>M</b>	<b>F</b>	<b>M</b>	<b>F</b>	<b>M</b>	<b>F</b>	<b>M</b>	<b>F</b>	<b>M</b>	<b>F</b>	<b>M</b>	<b>F</b>
<b>Alicante</b>	0.528 9421 1576 8463	0.46 0079 8403 19361	0.5479 041916 16767	0.46 8063 87225 5489	0.60 9780 4391 21757	0.46 0079 8403 19361	0.64 37125 7485 0299	0.47 7045 9081 83633	0.541 9161 6766 467	0.47 9041 9161 6766	0.45 0099 8003 9920	0.49 3013 9720 5588
<b>Benicasim</b>	0.731 5369 2614 7705	0.49 4011 9760 4790	0.6586 826347 30539	0.48 8023 9520 9580	0.75 0499 0019 9600	0.54 6906 1876 24751	0.76 8463 0738	0.558 88223 552	0.381 23752 495	0.528 9421 1576	0.24 9500 9980 039	0.51 8962 0758 4830
<b>Castellon</b>	0.66 6666 6666 6666	0.66 0678 6427 14571	0.7005 988023 9521	0.623 7524 9500 998	0.823 35329 34131 74	0.618 7624 7504 99	0.84 8303 39321 3573	0.593 81237 5249 501	0.61 9760 4790 4191	0.558 88223 5528 942	0.457 0858 2834 3313	0.51 8962 0758 4830

The definitions missing in the CSISP ontology for the statistical dataset are the units of the measured values and ontologies for statistical functions. The discussion for this issue is left to conclusion chapter, limitation and future work.

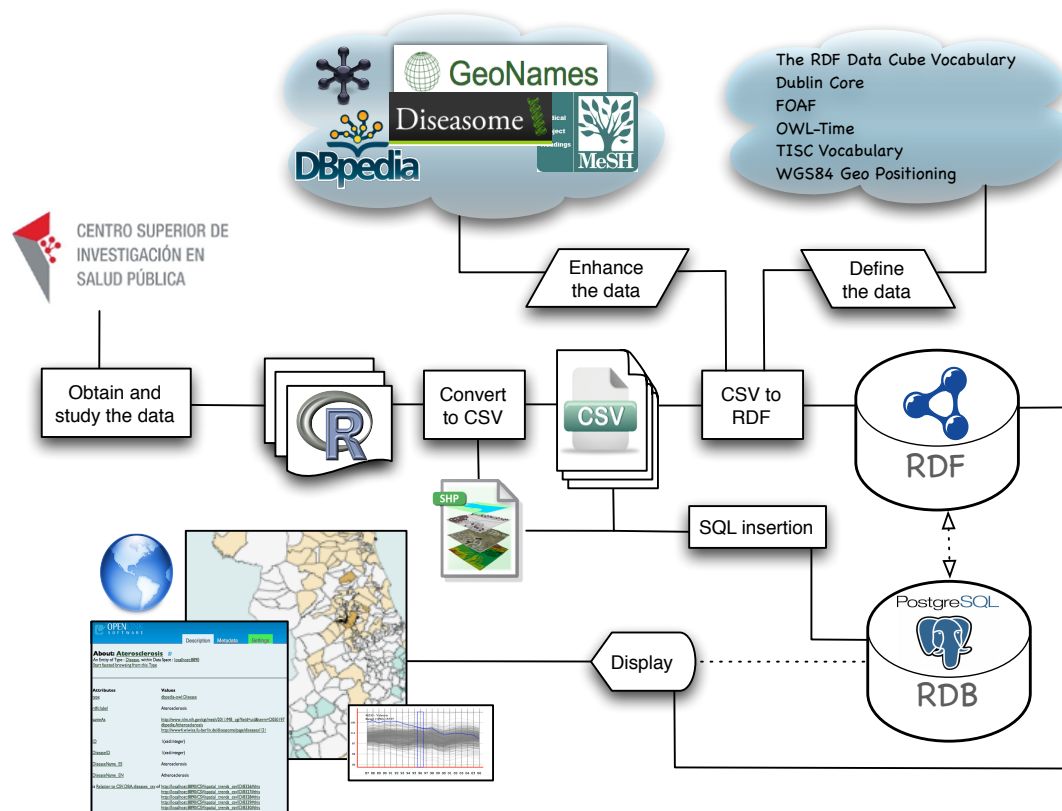
Overall steps taken in methodic background section and methodology of the data management chapter are depicted in the workflow diagram at Figure 7. Explanation of display and presentation option of the data are stated in Chapter IV with infrastructural capabilities of accessing the data in a distributed system via possible interfaces and API's under the system overview chapter. The sequence of the process are summarized as follows:

- Obtaining the data from CSISP and studying the stack of files in R-Data format.

<sup>48</sup> Especially time dimension is important in the case of infectious diseases like influenza, the study of their geographic distribution frequently involves examining the diffusion of the disease through space over a given period of *time* (spatio-temporal mapping and analysis) (Boulos K. , 2002).

- After the data is clarified in R-Data format is organized for converting to CSV with R scripts.
- Data is organized into topical domains in stack of CSV files which is more convenient for deciding to choose LOD datasets for enhancing the content of data and formal vocabularies to define the data. These datasets and vocabularies are shown in clouds in the diagram and depicted as inputs during the process of;
- CSV to RDF conversion: is accomplished with mapping language provided by RDF store.

Figure 7: Workflow Diagram of Methodology



- SHP files for the region of the use-case are also exported to CSV files as well as inserted to a spatial database with other CSV files.
- RDB to RDF Store connection for further replications is underlined but not accomplished due to licensing issues of the current RDF store.

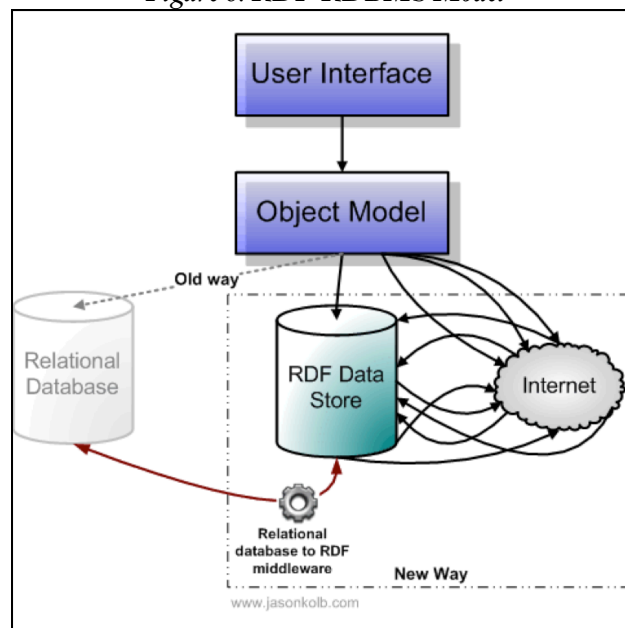
As a result of these methodological steps the data is structured in order to be served various channels, linkage between different domains of data sets is enriched and adding semantics enhanced the definition of the data.

## CHAPTER IV

# System Overview

Relational database and Linked Data integration carries an important value for enterprises; alleviation of heterogeneous data integration challenges and for public to discover linked data as well as create data mesh-ups. Since notable amount of web content is stored in relational databases there are several approaches (METAmorphoses<sup>34</sup>, D2RQ<sup>32</sup> and SquirrelRDF<sup>33</sup>) for generating semantic web data from relational databases (Svihla & Jelinek, 2007). Linked data generation from relational databases is sketched in Figure 3.

*Figure 8: RDF-RDBMS Model*



Source: [www.jasonkolb.com](http://www.jasonkolb.com)

An alternative and a robust tool for this purpose is Virtuoso with support of enterprise data sources (i.e. Oracle, SQL Server, Informix, Sybase, MySQL, PostgreSQL etc.). Virtuoso uses a SPARQL-based Meta Schema Language to provide RDBMS-to-RDF mapping functionality as well as automated generation

of linked data views of relational data. As stated earlier in methodic background section in chapter 2 Virtuoso is appointed as RDF store. Virtuoso Quad Patterns is the middleware translates relational data to RDF as Graph, Subject, Predicate, Object. Taking as an example of cities dataset; in the following listing initially formal vocabularies and ontologies are listed to define the data when it's mapped to RDF.

*Listing 5: Prefixes to define regions/cities dataset*

```
prefix CSISP: <http://localhost:8890/schemas/CSISP/>
prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix owl: <http://www.w3.org/2002/07/owl#>
prefix geo: <http://www.w3.org/2003/01/geo/wgs84_pos#>
prefix tisc: <http://observedchange.com/tisc/ns/#>
prefix gn: <http://www.geonames.org/ontology#>
prefix dbpedia-owl: <http://dbpedia.org/ontology/#>
```

A Virtual QuadStorage is created to store virtual RDF graphs from relational data in Virtuoso's relational data engine. Referenced tables with foreign keys (statistical tables of observations) are also included in order to link during mapping as in the following listing.

*Listing 6: RDF View Definition for regions dataset*

```
alter quad storage virtrdf:DefaultQuadStorage
from CSISP.DBA.regions as regions
from CSISP.DBA.spatial_trends as spatial_trends
where (^{spatial_trends.}.ComID = ^{regions}.gid)
from CSISP.DBA.spatio_temporal_trends as spatio_temporal_trends
where (^{spatio_temporal_trends.}.ComID = ^{regions_s.}.gid)
{ create CSISP:qm-regions as graph iri
("http://^{URIQADefaultHost}/CSISP#") {
```

RDF Entity Mapping for regions table is shown as the columns of relational data by matching with formal vocabularies to the values as literal, URI or scalar.

*Listing 7 : RDF Entity Mapping of regions dataset*

```
# Maps from columns of "CSISP.DBA.regions"
CSISP:regions a dbpedia-owl:PopulatedPlace ;
gn:PostalCode regions.gid as CSISP:regions-gid;
foaf:name regions.name as CSISP:regions-name ;
dbpedia-owl:PopulationTotal regions.population as CSISP:regions-
population;
owl:sameAs regions.geonamesURI as CSISP:regions-geonamesuri ;
rdfs:seeAlso regions.geonames_doc as CSISP:regions-geonames_doc;
tisc:areazsize regions.area as CSISP:regions-area ;
geo:lat regions.lat as CSISP:regions-lat ;
geo:long regions.lon as CSISP:regions-lon ;
```

Tables referenced in Listing 6 by joining their foreign keys with primary key in regions table are used to identify relations mapped into URI's with related associations in the following listing.

*Listing 8: Associations of regions dataset's entities with other entities*

```
# Maps from foreign-key relations of "CSISP.DBA.regions"
CSISP:regions_of CSISP:spatial_trends(spatial_trends.ID)
as CSISP:regions_spatial_trends_of ;CSISP:regions_of
CSISP:spatio_temporal_trends (spatio_temporal_trends.ID)
as CSISP:regions_spatio_temporal_trends_of .}};
```

CSV to RDF process shown at “Figure 7: Workflow Diagram of Methodology,” Relational Database to RDF shown with red arrow at “Figure 8: RDF-RDBMS Model” and RDF Wrapping of Structured and RDBMS data in Data Layer at “Figure 9: Multilayered Architecture” is basically the same process which is listed in from Listing 6 to Listing 8 with a built-in functionality of the established system. This process is repeated for all datasets until expected results are achieved. The process carries importance for increasing the quality of the data to be consumed in SDI and Linked Data applications coherently. SDI<sup>41</sup> initially mentioned at page 16 of this document as to be complementary with LOD and bootstrapping LOD with SDI said as an apprising strategy. SDI is widely used for different disciplines; environmental monitoring, emergency response, natural disasters and outbreaks, governmental resources and applications. SDIs are largely utilized by GIS domain. Linking SDI services with LOD by dereferencing the URIs of the data objects is a chance to spread out rich location data with cross-domain content. Overall the system is designated for reducing the boundaries between semantic web and geo web, support interoperability and extensibility for different platforms. Besides, building RDF schema attains added value of linked data over traditional way of storing data.

Subsequent to contribution and overview of the system that mentioned above this chapter elaborates the multi layer architecture of the system, singly for each layer and their elements with their functionalities/opportunities and introduce the user access methods with some examples.

## 4.1 Multi Layer Architecture

Multi layer architecture provides flexibility and reusability in which data



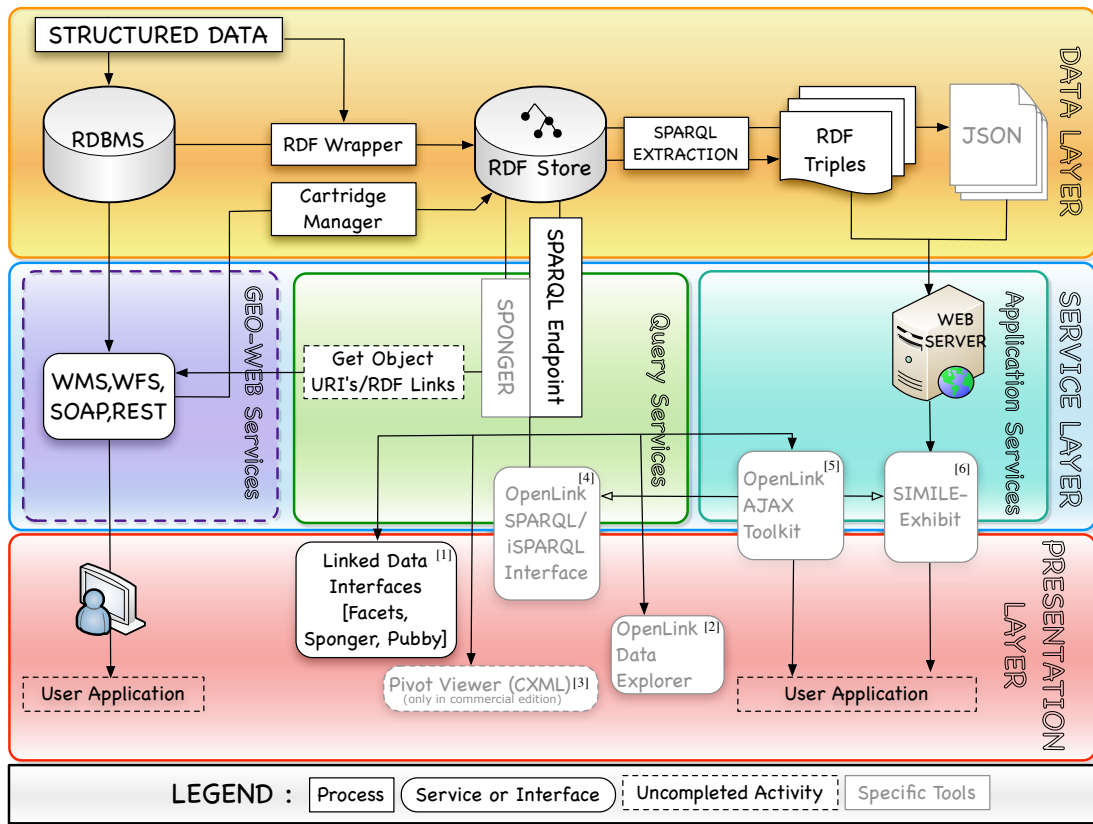
management, service processing and presentation are logically separate processes. Due to the scale and nature of the study multi layered architecture approach is preferred. By breaking up the system into layers, different layers can be developed sequentially and modified asynchronously without affecting the entire system. Multi layer architecture is composed of 3 main layers, which is a traditional client-server architecture depicted entirely in Figure 9: Multilayered Architecture. Data layer which consists of various data servers and formats that communicates to presentation layer through an intermediary layer composed by application, query and web services.

#### **4.1.1 Data Layer**

Data layer is which the information is stored and managed. Structured data refers to CSV (Comma Separated Values) files organized from the initial data received as R-data files in current case. CSV files, as a raw data format is compatible for insertion process to both, relational database and RDF store. PostgreSQL is chosen as a relational database due to the reason it's open source and supports convenient GIS data by PostGIS. Thus WMS (Web Mapping Services) and WFS (Web Feature Services) can be designated to the backend of PostgreSQL for transactions and creating intricate user interfaces. Another reason to establish a relational database is to keep the data which to be mapped into RDF in a sole environment instead of several CSV files. This approach also aids to track replication of data in both stores. Virtuoso offers more than a RDF store with promising capabilities for solving data silos problem. Especially for large scale enterprises, big companies and governmental organizations with various departments and disparate data sources in several formats, by delivering an unrivaled platform for real-time access and integration of relational databases, web services, and structured / semi-structured Web content / data resources.

RDF wrappers is a lightweight software components that is set up on top of an existing data source and exposes the data stored therein as RDF without affecting existing storage infrastructures (Haslhofer, Momeni, Schandl, & Zander, 2011). Virtuoso provides a built-in mechanism for RDF wrapping however system designated in Figure 9: Multilayered Architecture represents roughly a generic architecture independent of platforms and tools thus RDF wrapping is shown as a separate process between relational database and triple store which is a common practice for RDB to RDF mapping. Cartridge manager is a compound of entity extractor and ontology mapper.

Figure 9: Multilayered Architecture



Cartridge manager works as an RDFizer but not from relational data but from non-RDF web data sources like web services. Entity extractor performs initial data extraction and entities extracted from these sources mapped to RDF by using suitable ontologies. Extraction, mapping pipeline and flow is not explained here as this cartridge put forward as an option for RDF creation from other sources than relational data. Umbel<sup>49</sup> and Open Calais<sup>50</sup> are some tools for entity extraction and Snoggle<sup>51</sup> is an interoperable ontology mapper.

Other forms of data, which can be extracted from an RDF store, are in different triple formats and notations (i.e. N3, Turtle, RDF/XML, JSON etc.)<sup>52</sup> SPARQL endpoint is a protocol service that enables users (human or machine) to query RDF store via SPARQL language. It's important to establish the server,

49 <http://umbel.org>

50 <http://www.opencalais.com/>

51 <http://snoggle.semwebcentral.org/>

52 N3: <http://www.w3.org/DesignIssues/Notation3.html>

Turtle: <http://www.w3.org/TeamSubmission/turtle/>

RDF/XML: <http://www.w3.org/TR/rdf-syntax-grammar/>

JSON: <http://json-ld.org/spec/latest/>

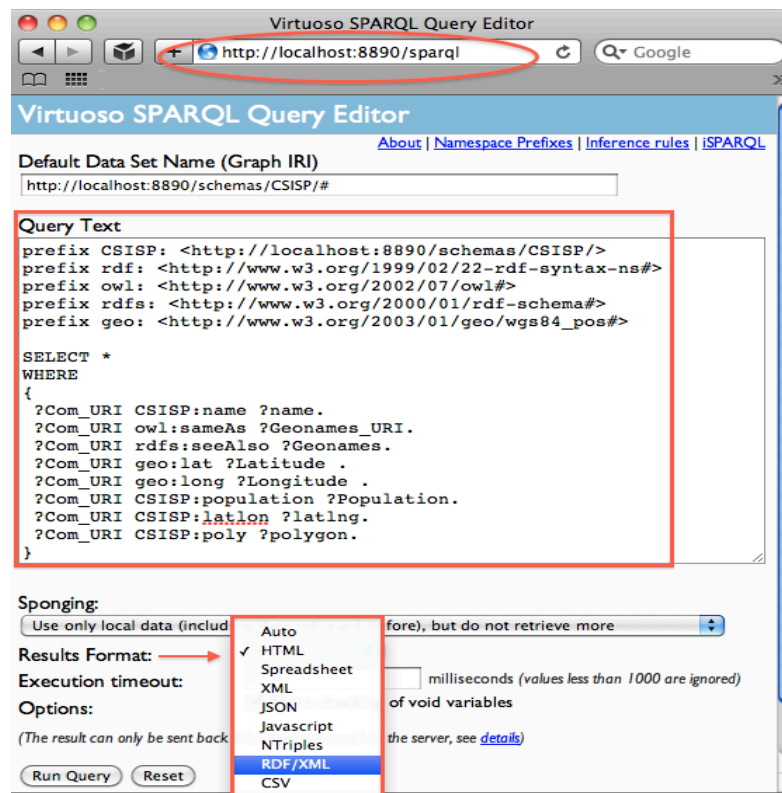
which facilitates to run a triple store with a SPARQL endpoint for remote queries and improved capabilities in order to fruitfully exploit the RDF data.

#### 4.1.2 Service Layer

Service layer is the intermediary layer, which basically controls the data access and bridges the client to the server via services by processing the commands of the clients. For this study case there are 3 segregated services; application, query and geo-web services. This structure leads also to define the roles of each services' segmentation within their context however this doesn't mean necessarily they are mutually exclusive. They are all web services from a general perspective though segmentation is about their functional context on the fly.

**Query Services:** SPARQL Query Language (W3C, 2008) an expressive language for formulating structured queries over RDF data sources. It defines a protocol for sending queries from clients to a SPARQL endpoint and for retrieving the requested results via the Web (Haslhofer, et al. 2011).

Figure 10: SPARQL Query Service Interface (4)

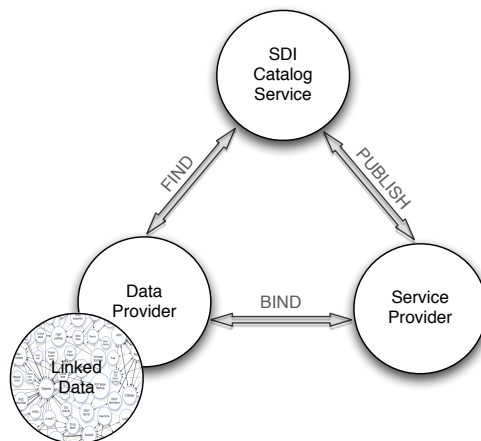


Source: *Running on localhost to be published online*

Virtuoso expose REST based SPARQL Web services supporting both GET and POST requests. An archetypal SPARQL endpoint interface can be seen in the following figure. Marked with red, consecutively with endpoint access address, a sample query and available data formats for the use case at *Figure 10: SPARQL Query Service Interface (4)*. Linked data client can communicate via SPONGER or SPARQL services. Sponger is not a generic tool therefore stated in the legend of Figure 9: Multilayered Architecture with other OpenLink Virtuoso specific tools. Sponger is comprised of cartridges coming with Virtuoso, which is highly customizable that custom cartridges can be developed (Virtuoso, 2009).

**Geo-Web Services:** This segment of the service layer hasn't been initiated yet however can be easily implemented for further application integrations and subtle visualizations. Therefore this segment is shown in Figure 4 as dashed box for this study. A feasible design of a user application would require map services for displaying the spatial trends of the mortal diseases through time. For serving this need, necessary object URIs can be get through query services of RDF store and merged with the designated service objects to display and browse through, on the user application. SDI is intended to design as a member of this segment with SOA-Service Oriented Architecture.

*Figure 11: SOA Operations of a Linked SDI*



Major functions in SOA include find, bind and publish are shown in Figure 11 above with key actors. In SOA based Linked SDI Data provider is enhanced to Linked Data sets from a constrained GIS user community. Linked data provides thematic attributions and enhanced information of (RDF) linked locations as mortality trends in Valencian community. Linked locations connected to SDI bridges Linked Data with Geo-Web.

**Application Services:** Application services includes server and tools to create custom web applications by consuming RDF data. Adjacent elements of service and presentation layer are AJAX toolkit and SIMILE exhibit libraries as specific tools due to access to services where they don't provide a complete end-user application but a user interface which anybody can develop custom web applications by accessing to services in this study.

### 4.1.3 Presentation Layer

Presentation layer is the final stage of implementation where user can interact with the services through various methods and interfaces. Linked data interfaces provide user nose-follow faceted browsing through the RDF data accessible from sparql endpoint. From a generic perspective linked data interfaces like Pubby<sup>37</sup> are supported by different RDF stores and Virtuoso comes integrated with Facets and Sponger interfaces. Following figure depicts an example of Facets interface depicted, which is fairly similar to Sponger.

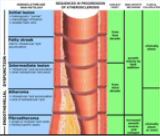
Figure 12: Faceted browsing of a disease between different sources (1)

**About: Atherosclerosis**

An Entity of Type : Thing, within Data Space : localhost:8890

[Start faceted browsing from this Type](#)

Atherosclerosis (also known as arteriosclerotic vascular disease or ASVD) is a condition in which an artery wall thickens as a result of the accumulation of fatty materials such as cholesterol.

Attributes	Values
type	Thing dbpedia-owl:Disease <a href="http://umbel.org/umbel/rc/AilmentCondition">http://umbel.org/umbel/rc/AilmentCondition</a>
rdfs:label	Atherosclerosis
rdfs:comment	Atherosclerosis (also known as arteriosclerotic vascular disease or ASVD) is a condition in which an artery wall th
sameAs	Atherosclerosis <a href="http://sw.opencyc.org/concept/Mx4r">http://sw.opencyc.org/concept/Mx4r</a> <a href="http://www4.wiwiwss.fu-berlin.de/sid">http://www4.wiwiwss.fu-berlin.de/sid</a> <a href="http://www4.wiwiwss.fu-berlin.de/sid">http://www4.wiwiwss.fu-berlin.de/sid</a> <a href="fbase:m/0lp66">fbase:m/0lp66</a> <a href="#">»more»</a>
name	Atherosclerosis
dbpprop:name	Atherosclerosis
depiction	

**About: Atherosclerosis**

An Entity of Type : CSV.DBA.diseases\_csv, within Data Space : localhost:8890

[Start faceted browsing from this Type](#)

Attributes	Values
type	CSV.DBA.diseases_csv dbpedia-owl:Disease
rdfs:label	Atherosclerosis
sameAs	<a href="http://www4.wiwiwss.fu-berlin.de/diseasome/page/diseases/121">http://www4.wiwiwss.fu-berlin.de/diseasome/page/diseases/121</a> dbpedia:Atherosclerosis <a href="http://www.nlm.nih.gov/cgi/mesh/2011/MB_cgi?field=uid&amp;term=D050197">http://www.nlm.nih.gov/cgi/mesh/2011/MB_cgi?field=uid&amp;term=D050197</a>
ID	1 (xsd:integer)
DiseaseID	1 (xsd:integer)
DiseaseName_ES	Atherosclerosis
DiseaseName_EN	Atherosclerosis

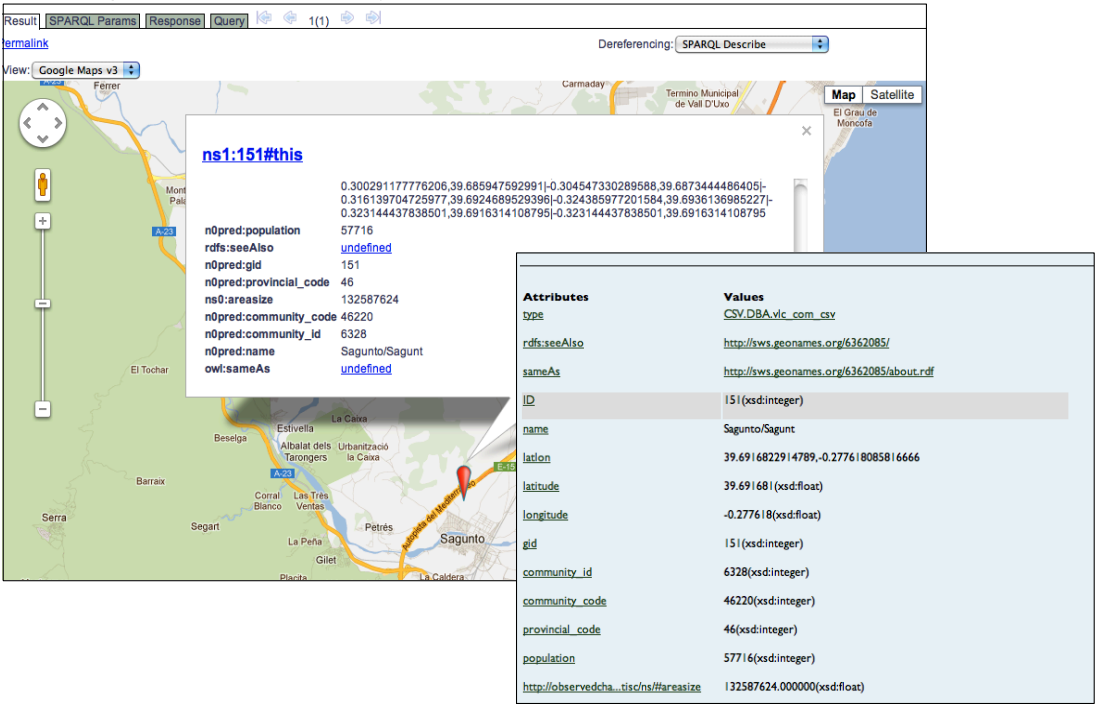
Wide ranges of options are shown for creating user applications in presentation layer of Figure 9: Multilayered Architecture. Variety of the tools also allows lay people to build their own demos and support public usage and accurate

interpretation as long as transparency and access of data is granted from authorities. As it's stated in the first chapter technical constraints and issues are focus of the study rather than privacy, security issues. Specific applications and sample demos for the use case study are discussed in the following section.

## 4.2 User-Access Methods

Users, developers, researchers, anyone who would like to consume this data can access through various interfaces. SPARQL endpoint serves the data like a proxy service to Linked Data frontends. SPARQL endpoint can be accessed directly through and HTML browser and complex queries can be sent to receive comprehensive replies as well as SPARQL endpoint can be accessed from third party applications or frameworks for statistical analysis and developing applications with map, timeline and graphic visualizations. (i.e.: R-Project SPARQL package, Pivotviewer, Openlink AJAX Toolkit, SIMILE-EXHIBIT Framework.)

Figure 13: Representation of a spatial linked data with OpenLink Data Explorer (2)



**(1) Linked Data Interfaces** provides a catalog like interface where people can browse the data between different sources through URI's. An example snapshot of a linked data interface is already shown Figure 12. For

searching particular data objects in the dataset users can access this information via SPARQL query service interfaces or API's, which access to endpoint and make a manual query. This approach may not so practical for basic users. Virtuoso Facets provides a Precision Search & Find service for text search, label of the entity search and search through URIs. Users can also navigate between entity relations

**(2) OpenLink Data Explorer<sup>53</sup>** is an extension to the supported browsers (Firefox, Safari, Google Chrome) with an additional browser support for viewing data sources associated to web pages. ODE uses a Model-View-Controller architectural pattern. Data display and presentation in view is shown in Figure 13: Representation of a spatial linked data with OpenLink Data Explorer (2) with the functionality of where it is on the map.

**(3) Pivot Viewer<sup>54</sup>** is a Silverlight<sup>55</sup> control so it also requires a Silverlight extension to the browser. Data collection is supposed to be Collection XML - CXML in which the format is supported to be extracted with Virtuoso. Pivot viewer provides a fancy interface by viewing linked data in CXML format an alternative to OpenLink Data Explorer though only supported in commercial edition. Both OpenLink Data Explorer and Pivot Viewer access to sparql endpoint of the triple store and according to the level of flexibility and integrity of the data, displays the result with map, timeline and graphs.

**(4) iSPARQL<sup>56</sup> Interface** is built by using OpenLink AJAX Toolkit . Interactive SPARQL – iSPARQL framework is an extension of SPARQL. It allows querying by triple patterns, conjunctions, disjunctions and optional patterns by enriching the traditional grammar syntax of SPARQL. It creates virtual triples that are not in the matching ontology but used to establish virtual relations. The relations also can be created with a graph drawing interface, which makes easier to establish relations. By using iSPARQL framework users can access to SPARQL endpoint of the established ontology for the study and create advanced relations with other datasets.

**(5) OpenLink AJAX Toolkit<sup>57</sup>** is a JavaScript-JS based toolkit for browser independent application development. In this users can access to data via

---

<sup>53</sup> <http://ode.openlinksw.com/>

<sup>54</sup> <http://www.openlinksw.com/PivotViewer>

<sup>55</sup> [www.silverlight.net](http://www.silverlight.net)

<sup>56</sup> <http://oat.openlinksw.com/ispardl/index.html>

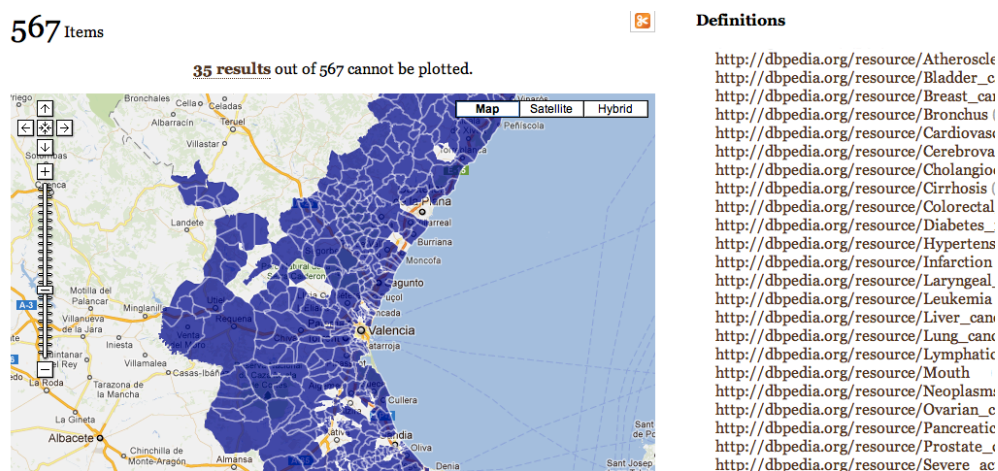
<sup>57</sup> <http://oat.openlinksw.com/>



various channels (SPARQL, SQL, REST, SOAP) according to availability. Web form designer and database designer is provided to control the widgets as well as API can be used manually through other application development environments.

**(6) SIMILE-Exhibit<sup>58</sup>** is an open source software platform for publishing linked data in JSON and RDF/XML formats. In the Figure 9: Multilayered Architecture in which the platform is receiving data shown only in JSON format was the only format, which is tested. Exhibit lets to even basic users create web pages with advanced search and filtering, faceted browsing functionalities, presentation with interactive maps, timelines and graphs. Simile Exhibit receives its own JSON format different from the one extracted by triple store virtuoso. Therefore a SparqlProxy<sup>59</sup> can be used to query a sparql endpoint and convert to different formats requested by visualization APIs. A proxy request is written to sparql endpoint of virtuoso to receive data in Exhibit JSON format. The result is shown in the following figure. To organize the facets and filter the data an advanced sparql xml to exhibit json bindings required to be defined.

*Figure 14: SIMILE Exhibit data representation from Virtuoso*



From a systematic perspective how can semantic technologies are used to integrate and present data for wide range of users are evaluated. Available channels are outlined in this chapter; the general overview of the system and possible contribution sources are drawn and mentioned in conclusion chapter.

<sup>58</sup> <http://www.simile-widgets.org/exhibit/>

<sup>59</sup> SparqlProxy is a web service that executes SPARQL query and convert the query results into formats needed by visualization APIs (TWC LOGD, 2010)



# Conclusions and Future Work

This chapter summarizes the outcomes and results obtained during the research for this thesis. Section 5.1 discusses the prospected contribution with respect to the problems phrased in chapter 1. Limitations and problems encountered during progress are discussed Section 5.2 with alternative methods and lessons learned with an outlook on future work and potential application areas in the final section 5.3.

## 5.1 Summary

The goal of this thesis was to a novel approach to convey complex health related information with a robust infrastructure that can access various sources. Proposed method was achieving interoperability in syntactic and semantic level by an infrastructure that interacts between RDB data to RDF data and creates a Linked Spatial Data Infrastructure across various domains. For this reason, database and Linked Data integration carries an important value for enterprises; alleviation of heterogeneous data integration challenges and for public to discover linked data as well as create data mash-ups. Overall interoperability and extensibility assured as RDF allows integration of data in a platform independent manner. Added value of Linked Data is achieved by building RDF schema over traditional way of storing data. Absence of semantic relation for logical data model is compensated by RDF based conceptual model that utilizes the semantic expressivity of RDF, RDFS, OWL, FOAF and domain related ontologies.

Statistical data shaped in a form, which represents more than just numbers but also advanced meta-data with dimensions of the observation regarding to space and time by ontological approach. Spatial health data for mortality rates carried

over into a system with an ontological based approach that can attend to several visualization technologies and integration with web of data for complex queries. Our approach for the model retains, relational databases and web services for application integration, visualization and deploy linked data for disparate data meshing, discovery and drill down analysis. These issues are discussed in the following sections.

## 5.2 Limitations and Lessons Learned

In this section limitations and obstacles faced in processes based on the workflow diagram at Figure 7 that carried on during the study are outlined and lessons learned are listed.

### 5.2.1 Limitations

**L.1. Lack of tools and best practices:** Despite there exists a huge literature behind the conceptual domains as stated in chapter 2, there is lack of connecting these domains in terms of tools and best practices between semantic web and geo-web. GeoSPARQL as the major bridge between linked data and geographical data; all triple stores do not enable it. Even though there are approaches with 3<sup>rd</sup> party tools, scripts and studies for complex geo data mapping there is a need of Geo2RDF converters. Spatio-temporal data publishing orchestrated with space, time vocabularies and statistical data publishing with RDF cube vocabulary, which is not a stand-alone dataset, but communicable with other linked data sets are lacking best practices.

**L.2. Semantics of Mappings and Linkages:** "For instance, who has attempted to outline the process and components of even a relatively small enterprise has experienced the brain-cramps that can come with complex ontology" says David Koepsell, Center for Commercial Ontology. As we stated in introduction chapter, problem section complex ontologies for large domains of objects are daunting. During the designation process of an ontology to enrich the content of the data and bring the semantics when defining the data brain-cramps are inevitable. Based on the use case data converted to RDF there are remained definitions should be healed with elaborated predicates and relations. Taking as an example the statistical data set; RDF Cube vocabulary does not yet represent the semantics of what actually the statistical data and functions describes. By adding data object descriptions with dimensions, measures,

attributes and observations reuse the existing definitions in the instance level or on the schema level. Succinctly translation of the statistical data to linked data use simple transformation approaches. Semantics of the functions, how they have been derived, units of measures, what real world concepts they represent as data points are lacking in statistical RDF data. Another erroneous assumption is the usage of `owl:sameAs` predicate which is ubiquitous in interlinking datasets. Predicates for interlinking local data to be published with the data from linked open data cloud may not be always accurate.

### 5.2.2 Lessons Learned

**L.L.1. Better statistical definitions and elaborated linkage predicates:** Grounding the aggregated statistical values and their functions to derive these values requires advanced ontologies specified in mathematical functions and defining units of measures like The QUDT<sup>60</sup> ontology (Quantities, Units, Dimensions and Data Types), SWEET<sup>61</sup> ontology (Semantic Web for Earth and Environmental Terminology) for representing mathematical concepts, (including arithmetic operators and statistical functions) or OpenMath<sup>62</sup> and such. Ambiguity of linking data with `owl:sameAs` needs to be reviewed. The question should be asked if interlinked URI's with this predicate are really the same thing or same thing as but different context. Thus more accurate linkage options should be considered like `skos:narrowMatch`, `skos:closeMatch`, `rdfs:seeAlso`. It worth to consider using Silk<sup>63</sup> discovery framework might help for interlinking process rather than semi-automated approach held.

**L.L.2. Collaboration with domain experts:** Anyone interested in the geography of disease will need a good understanding of the basics of epidemiology, or at least of health statistics. Therefore the fundamental premise of linking health data should be studied and according to the addressing fields aside from health, geography and statistics but also environmental studies (concerning about exposure/hazard outcomes) experts need to be included in the progress.

**L.L.3. Fundamental premise of linking health data:** The very first argument should be considered if the data is adequate and appropriate for

---

<sup>60</sup> QUDT- <http://www.qudt.org>

<sup>61</sup> SWEET- <http://sweet.jpl.nasa.gov/ontology/>

<sup>62</sup> OpenMath - <http://www.openmath.org/>

<sup>63</sup> Silk - <http://www4.wiwiiss.fu-berlin.de/bizer/silk/>

addressing a possible issue and promise to fruitfully exploit information from other datasets. Framework is useful to follow which is offered by Thacker et al. (1996) elucidate the steps whereby an environmental agent moves through the environment (hazard), enters a person (exposure), and produces an effect (health outcome).

**L.L.4. Smaller scale projects for a start:** Considering the complexity of ontology designation, data management and publishing of cross-domain data; it's clear that starting with smaller scale datasets would help optimizing the outcomes of the study. Parsing the study between domains, applying each single domain with an ultimate structure would be beneficiary to understand the concepts achieve premises stated above in a longer term but more efficiently.

## 5.3 Future Work

This section rounds the thesis off with directions for future implementations and studies.

### 5.3.1 Utilizing OGC Services

Remained part of the infrastructure to implement Geo-Web services is one of the directions of the study for establishing linked SDI. WFS – Web Feature Services as a part of SDI implementation could connect Linked Data and OGC. WFS request can take URI and return results in RDF as well as return results mapped directly into given domain models. Also increasing the means of usage for GeoSPARQL standard of OGC is another intersection point with OGC services. Leveraging semantics with sensor information is an idea on the fly yet but possible with OGC Sensor Data from environmental and health monitors. For example heat-related mortality on heat islands is directly related with health data and temperature data can be received from sensor web and overlaid to geographical data. Similar relation can be exemplified between ozone levels and asthma

### 5.3.2 Linking with Environmental Data

Environmental profiles of the location affect the prevalence of the diseases as stated in initial chapter. For linking the health data with environmental data, relation between the existing diseases and environment should be studied.

Etiology of mortality and morbidity based on environment can be caused from following examples of environmental data and sources; toxic release inventories, air pollutant data like ozone, sulfur dioxide, CO<sub>2</sub>, pesticide exposures, water reservoirs and safe drinking water information sources (Mather, et al., 2004). After studying the relations between health and environmental data geographic scale of the study needs to be specified. Following this pattern aggregated environmental hazard or exposure data can be correlated with aggregated health data for each unit of observation by geographic boundaries and extents covering a contamination region. With the proposed Linked SDI architecture INSPIRE<sup>64</sup> – Infrastructure for Spatial Information in the European Community and GEOSS<sup>65</sup> – Group on Earth Observation System of Systems are promising vendors by their background and studies.

### 5.3.3 Enriching the Statistics

For analyzing and linking disparate data in a scientifically manner complex relationships between health data and health related data requires attentive consideration of the statistical models. Etiologic factors of mortality and morbidity are not constrained with geography and its environmental factors. Age, race, sex, deprivation, scale/ratio of correlation areas are other controlling covariates. Linking cross-domain data in a machine understandable way doesn't solve the whole issue but ease to access related information. For reasoning and analyses statistical methods should be enriched.

A step further for statistical computing in machine consumable way is to use SPARQL package in R-project to handle statistical health data. Thus the objective of the study could be exploited by reusing the channels of information via established infrastructure for creating a next generation health representation reasoning tools and visualizations.

---

<sup>64</sup> <http://inspire.jrc.ec.europa.eu/>

<sup>65</sup> <http://www.earthobservations.org/geoss.shtml>

# Bibliography

- Abadi, D. J., Marcus, A., Madden, S. R., & Hollenbach, K. (2007). Scalable Semantic Web Data Management Using Vertical Partitioning. *In Proceedings of the 33rd international conference on Very large data bases - VLDB '07*.
- AEMET. (2010). *AemetLinkedData.es*. (O. E. (OEG), Producer) Retrieved February 2012, from [http://aemet.linkeddata.es/technology\\_en.html](http://aemet.linkeddata.es/technology_en.html)
- Battle, R., & Kolas, D. (2011). Enabling the Geospatial Semantic Web. *Semantic Web Journal*.
- Belleau, M., Tourigny, N., & Rigault, P. M. (2008). Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. *Journal of Biomedical Informatics*, 41(15):706-16, 2008. (41), 706-16.
- Berners-Lee, T. (2006, July 27). *Linked Data-Design Issues*. Retrieved February 2012, from <http://www.w3.org/DesignIssues/LinkedData.html>
- Bizer, C., Cyganiak, R., Garbers, J., Maresch, O., & Becker, C. (2009). *The D2RQ Platform v0.7 - Treating Non-RDF Relational Databases as Virtual RDF Graphs*. Retrieved February 2012, from <http://www4.wiwiwiss.fu-berlin.de/bizer/d2rq/spec/#specification>
- Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems*.
- Bizer, C., Heath, T., Ayers, D., & Yves, R. (2001). Interlinking Open Data on the Web.
- Bishr, Y. (1998). Overcoming the semantic and other barriers to GIS interoperability. *International Journal of Geographical Information Science* (12), 299-314.
- Blakeley, C. (2007). *"RDF Views of SQL Data (Declarative SQL Schema to RDF Mapping)"*. OpenLink Software.
- Boulos, K. M. (2003). Location-based health information services: a new paradigm in personalised information delivery. *International Journal of Health Geographics*, 2(2).
- Boulos, K. M. (2004). Towards evidence-based, GIS-driven national spatial health information infrastructure and surveillance services in the United Kingdom. *International Journal of Health Geographics*, 3(1).
- Boulos, K. (2002, October). *Medical Geography*. (K. Boulos, Editor, K. Boulos, Producer, & School of Informatics City University, London, UK) Retrieved February 2012, from Health Geomatics: <http://health-geomatics.co.nr/>
- Brownstein, J., Freifeld, C., Reis, B., & Mandl, K. (2007). HealthMap: Internet-based emerging infectious disease intelligence. *National Academy of Science*, 183-204.
- Cyganiak, R., Reynolds, D., & Tennison, J. (2010, July 14). *The RDF Data Cube vocabulary*. Retrieved February 2012, from <http://publishing-statistical-data.googlecode.com/svn/trunk/specs/src/main/html/cube.html>
- Cochrane, R. G., & Galperin, Y. M. (2010). The 2010 Nucleic Acids Research Database Issue and online Database Collection: a community of data resources. *Nucl. Acids Res* (38), D1-D4.
- EnAKTing. (2009). *EnAKTing Forging the Web of Linked Data*. (N. Shadbolt, T. Berners-Lee, W. Hall, & N. Gibbins, Editors) Retrieved February 2012, from <http://www.enakt.org/>
- ESRI. (1998). *ESRI Shapefile Technical Description*. White Paper, ESRI.
- Erling, O., & Mikhailov, I. (2007). RDF Support in the Virtuoso DBMS. In S. Auer, C. Bizer, C. Müller, & A. V. Zhdanova (Ed.), *Proceedings of the 1st Conference on Social Semantic Web CSSW*, 113, pp. 59-68.
- Daconta, M. C., Obrst, L. J., & Smith, K. T. (2003). Understanding the Resource Description Framework. In M. C. Daconta, L. J. Obrst, & K. T. Smith, *The Semantic Web* (Vol. 5, pp. 85-119). Indianapolis, USA: Wiley.
- Guarino, N. (1998). Formal Ontology and Information Systems. In N. Guarino (Ed.), *1st Int. Conf. on Formal Ontology in Information Systems* (pp. 3-15). IOS Press.
- Goodchild, M. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69, 211-221.
- Goodchild, M. (2011). Challenges in geographical information science. *Proc. R. Soc.* (467), 2431-2443.
- Gruber, T. (1993). A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition* (5), 199-220.
- IMD. (2010). *Index of Multiple Deprivation Linked Data*. Retrieved February 2012, from Open Data Communities. : <http://opendatacommunities.org>
- Hausenblas, M., Halb, W., Raimond, Y., Feigenbaum, L., & Ayers, D. (2009). SCOVO : Using Statistics on the Web of Data. *The Semantic Web Research and Applications* (5554), 708-722.

<p>Haslhofer, B., Momeni, E., Schandl, B., &amp; Zander, S. (2011). <i>Europeana RDF Store Report</i>. University of Vienna, Research Group Multimedia Information Systems Faculty of Computer Science.</p> <p>Hastings, J. (2008). Automated Conflation of Digital Gazetteer Data. <i>International Journal of Geographical Information Science</i> (22), 1109-1127.</p> <p>HCLSIG. (2009, August ). <i>Linking Open Drug Data</i> . Retrieved February 2012, from HCLSIG/LODD: <a href="http://www.w3.org/wiki/HCLSIG/LODD/Data">http://www.w3.org/wiki/HCLSIG/LODD/Data</a></p> <p>Heath, T., &amp; Bizer, C. (2011). <i>Linked Data: Evolving Web into a Global Data Space</i> (1st Edition ed.). (J. Hendler, &amp; F. Harmelen, Eds.) Morgan&amp;Clayppol Publishers.</p> <p>HGP. (2003). <i>HGP-Human Genome Project</i>. Retrieved February 2012, from Human Genome Project Information: <a href="http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml">http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml</a></p> <p>Jacobs, I., &amp; Walsh, N. (2004). <i>Architecture of the World Wide Web</i> . Retrieved from W3C: <a href="http://www.w3.org/TR/webarch/">http://www.w3.org/TR/webarch/</a></p> <p>Jentzsch, A., Hassanzadeh, O., Bizer, C., Andersson, B., &amp; Stephens, S. Enabling Tailored Therapeutics with Linked Data. In <i>Proceedings of the 2nd Workshop on Linked Data on the Web</i>, 2009.</p> <p>Johnson, S. (2006). <i>The Ghost Map: The Story of London's Most Terrifying Epidemic and How it Changed Science, Cities and the Modern World</i>. Riverhead Books.</p> <p>Kuhn, W. (2005). Geospatial semantics: Why, of what and how? . <i>Journal on data semantics III</i>. LNCS 3534. , 1-24.</p> <p>Kuhn, W. (2005, March 15). Introduction to Spatial Data Infrastructures. Muenster.</p> <p>Keller, M. (2009). Use of Unstructured Event-Based Reports for Global Infectious Disease Surveillance. <i>Emerging Infectious Diseases</i> , 15 (5), 689-695.</p> <p>Knuth, D. E. (1998). Sorting and searching. In <i>The art of computer programming</i> (2nd Edition ed., Vol. III, pp. 395-395). Addison-Wesley.</p> <p>NOAD. (2005). <i>The New Oxford American Dictionary</i> (2nd Edition ed.). (E. McKean, Ed.) Oxford University Press.</p> <p>Mather, F. J., Ellis, W. L., Langlois, E. C., Shorter, C. F., Swalm, C. M., Shaffer, J., et al. (2004). Statistical Methods for Linking Health, Exposure, and Hazards. <i>Environ Health Perspect</i> , 112, 1440-1445.</p> <p>McCarthy, M., Abecasis, G., &amp; Cardon, L. (2008). Genome- wide association studies for complex traits: consensus, uncertainty and challenges. <i>Nat Rev Genet</i> , 9, 356-69.</p> <p>MeSH. (2012). <i>National Library of Medicine - Medical Subject Headings</i> . Retrieved February 2012, from <a href="http://www.nlm.nih.gov/mesh/meshhome.html">http://www.nlm.nih.gov/mesh/meshhome.html</a></p> <p>Momtchev, V. (2009). <i>Expanding the Pathway and Interaction Knowledge in Linked Life Data</i>. LarKC The Large Knowledge Collider: a platform for</p>	<p>large scale integrated reasoning and Web-search, Sofia.</p> <p>PubMed. (2010). <i>PubMed Health</i>. Retrieved February 2012, from <a href="http://www.ncbi.nlm.nih.gov/pubmedhealth/PMH0002267/">http://www.ncbi.nlm.nih.gov/pubmedhealth/PMH0002267/</a></p> <p>Pew Environmental Health Commission. (2000). <i>America's Environmental Health Gap: Why the Country Needs a Nationwide Health Tracking Network</i>. MD. Baltimore: Technical Report.</p> <p>Pehle, T. (2010, June). <i>Bootstrapping the Web of Linked Locations</i>. Retrieved February 2012, from <a href="http://ogcnetwork.org">ogcnetwork</a>.</p> <p>Pehle, T. (2011, June). Geographic Information in Government Linked Data - W3C Government Linked Data F2F.</p> <p>Svihla, M., &amp; Jelinek, I. (2007). Benchmarking RDF Production Tools. . <i>18th International Conference on Database and Expert Systems Applications DEXA 2007</i> (pp. 700-709). Springer.</p> <p>Smart, P., Jones, C., &amp; Twaroch, F. (2010). Multi-source Toponym Data Integration and Mediation for a Meta-Gazetteer Service. <i>GIScience</i> , 234-248.</p> <p>Stolze, K. (2003). SQL/MM Spatial: The Standard to Manage Spatial Data in Relational Database Systems. <i>Datenbanksysteme in Büro, Technik und Wissenschaft</i> .</p> <p>Russell, S. J., &amp; Norvig, P. (2003). <i>Artificial Intelligence: A Modern Approach</i> (2nd Edition ed.). New Jersey: Prentice Hall.</p> <p>Revelytix, Inc. . (2010). <i>Triple Store Evaluation Analysis Report</i>. Maryland.</p> <p>TWC LOGD . (2010). <i>Linking Open Government Data</i> . Retrieved February 2012, from Tetherless World Constellation: <a href="http://logd.tw.rpi.edu/technology/SparqlProxy">http://logd.tw.rpi.edu/technology/SparqlProxy</a></p> <p>Tauberer, J. (2005, February 3). <i>Using RDF for Distributed Information</i> . Retrieved February 2012, from <a href="http://govtrack.us">govtrack.us</a>: <a href="http://www.govtrack.us/articles/20050302rdf.xpd">http://www.govtrack.us/articles/20050302rdf.xpd</a></p> <p>Thacker, S., Stroup, D., Parrish, R., &amp; Anderson, H. (1996). Surveillance in environmental public health: issues, systems and sources . <i>Journal of Public Health</i> (86), 633-638.</p> <p>The World Bank. (2010, May 21). <i>Data World Bank</i>. Retrieved February 2012, from <a href="http://data.worldbank.org/news/rosling-noveck-event-may-21-2010">http://data.worldbank.org/news/rosling-noveck-event-may-21-2010</a></p> <p>Valle, E., Qasim, H. M., &amp; Celino, I. (2010). Amalgamated System for Treating Non-RDF (Relational &amp; Spatial) Data Stores as Virtual RDF Graph. <i>ISPRS - International Society for Photogrammetry and Remote Sensing, XXXVIII</i>, pp. 4-13.</p> <p>Virtuoso. (2009, September). <i>Virtuoso Programmer's Guide</i> . Retrieved February 2012, from RDF Middleware ("Sponger"): <a href="http://virtuoso.openlinksw.com/dataspace/dav/wiki/Main/VirtSpongerCartridgeProgrammersGuide">http://virtuoso.openlinksw.com/dataspace/dav/wiki/Main/VirtSpongerCartridgeProgrammersGuide</a></p>
---	--

Virtuoso. (2010). *Mapping Relational Data to RDF with Virtuoso's RDF Views*. Retrieved February 2012, from <http://virtuoso.openlinksw.com/whitepapers/relational%20rdf%20views%20mapping.html>

Vrandečić, D., Lange, C., Hausenblas, M., Bao, J., & Ding, L. (2010). Semantics of Governmental Statistics Data. *proceedings of the WebSci10: Extending the Frontiers of Society On-Line*. Raleigh, NC: WebScience Trust.

W3C. (2001). *W3C Semantic Web*. Retrieved February 2012, from <http://www.w3.org/2001/sw/>

W3C. (2008). *SPARQL Query Language for RDF*. (W. W. Consortium, Producer) Retrieved February 2012, from W3C Semantic Web Activity – RDF Data Access Working Group.

Walport, M., & Brest, P. (2011). Sharing research data to improve public health. *Lancet* (377), 537-9.

Zaveri, A. J. (2011, June). *AKSW: Projects / Stats2RDF*. Retrieved February 2012, from Representing multi-dimensional statistical data as RDF using the RDF Data Cube Vocabulary: <http://aksw.org/Projects/Stats2RDF>

Zaveri, A., Pietrobon, R., Ermilov, T., Martin, M., & Heino, N. (2010). *Evaluating the disparity between active areas of biomedical research and the global burden of disease employing Linked Data and data-driven discovery*. . Tuberculosis IMISE Report.