

MASTERS PROGRAM IN



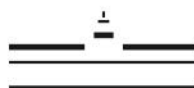
GEOSPATIAL TECHNOLOGIES

**Linked Data based Health Information Representation,
Visualization and Retrieval System on the Semantic Web**

Prepared by:

Binyam Chakilu Tilahun

Dissertation submitted in partial fulfilment of the requirements
for the Degree of *Master of Science in Geospatial Technologies*



WESTFÄLISCHE
WILHELMS-UNIVERSITÄT
MÜNSTER

**Linked Data based Health Information Representation,
Visualization and Retrieval System on the Semantic Web**

Dissertation supervised by

Dr. Tomi Kauppinen (PhD)

Dissertation Co-supervised by

Dr. Carsten Keßler (PhD)

Prof. Marco Painho (PhD)

January 2013

Münster, Germany

Declaration

I hereby, certify that I have written this thesis independently with the guidance of my supervisors and with no other tools than the specified. Data was extracted primarily from WHO global health observatory dataset and missing data was complemented by CDC and UNAIDS datasets with the given right of implementing and publishing in Linked Open Data. Sources used are referenced in the bibliography.

February 20, 2013
Muenster, Germany

Acknowledgment

I would like to express my deepest gratitude to my Supervisor Dr. Tomi Kauppinen for all his support, valuable comments, prompt replies, and discussions with his innovative ideas that hugely inspire and support me to finish this thesis on time. Linked Open Data is completely a new science for me, thank you for inviting me to the conference in Paris that helped me shape the idea of this thesis.

I would also like to extend my gratitude to my co-supervisors- Dr. Carsten Kießler, and Professor Marco Painho for their comments, ideas and corrections. I would also like to thank Prof. Werner Kuhn, Prof. Edzer Pebesma and prof. Chris Kray for their monthly follow up, encouragement, and vital comments. I am also thankful to Dori and Karsten, for being a click away in all aspect of help we need. You guys are families, thank you!

I would also like to thank Data geeks in WHO, UNAIDS and other organizations who make the health data I use for testing the system, available online for free. In the future, I hope those guys will represent their data in Linked Open Data and peoples can re-use and play with it with out too much effort and time investment.

I am grateful to European Commission for the dual opportunity they gave me by funding this study. It was a big opportunity and blessing to live and learn with different students from different countries, which have different culture, language and way of life. I am thankful for all colleagues in Spain and Germany especially, Talaksew for making my life in Münster memorable. Our discussions, not only academics but also social and political, at library coffee machine, makes my time wonderful. I am happy we are Friends and we will be.

Last but obviously not least, I would like to acknowledge my beloved wife, Rosina, for sharing her life with me, for motivating me to do this degree and take care of our little angel, Yordanos, in my absence from home for more than a year. Thank you our angel for making us feel responsible at this age of our life. I love you both!!!

Abstract

To better facilitate health information dissemination, using flexible ways to represent, query and visualize health data becomes increasingly important. Semantic Web technologies, which provide a common framework by allowing data to be shared and reused between applications, can be applied to the management of health data. Linked open data - a new semantic web standard to publish and link heterogonous data- allows not only human, but also machine to brows data in unlimited way.

Through a use case of world health organization HIV data of sub Saharan Africa - which is severely affected by HIV epidemic, this thesis built a linked data based health information representation, querying and visualization system. All the data was represented with RDF, by interlinking it with other related datasets, which are already on the cloud. Over all, the system have more than 21,000 triples with a SPARQL endpoint; where users can download and use the data and – a SPARQL query interface where users can put different type of query and retrieve the result. Additionally, It has also a visualization interface where users can visualize the SPARQL result with a tool of their preference. For users who are not familiar with SPARQL queries, they can use the linked data search engine interface to search and browse the data.

From this system we can depict that current linked open data technologies have a big potential to represent heterogonous health data in a flexible and reusable manner and they can serve in intelligent queries, which can support decision-making. However, in order to get the best from these technologies, improvements are needed both at the level of triple stores performance and domain-specific ontological vocabularies.

Keywords: Linked Open Data, semantic web, ontology, health information, HIV, WHO

Acronym

LOHD	-	Linked Open Health Data
LOD	-	Linked Open Data
WHO	-	World Health Organization
OWL	-	Ontology Web Language
RDF	-	Resource Description Framework
SWRL	-	Semantic Web Rule Language
UMLS	-	Unified Medical Language System
UML	-	Unified Modeling Language
WPS	-	Web Processing Service
W3C	-	World Wide Web Consortium
XML	-	Extensible Markup Language
URI	-	Unified Resource Identifier
URL	-	Unified Resource Locator
RDF	-	Resource Description Framework
HTTP	-	Hypertext Transfer Protocol
WWW	-	World Wide Web
XML	-	Extensible Markup Language
DBMS	-	Database Management System
OWL	-	Ontology Web Language
GIS	-	Geographic Information Systems
W3C	-	World Wide Web Consortium
FOAF	-	Friend of a friend
SPARQL	-	SPARQL protocol and query language

1 Table of Contents

Abstract	v
Acronym.....	vi
Table of Contents	vii
1. Introduction	1
1.1. Motivation	3
1.2. Problem statement.....	4
1.3. Solution Approach.....	5
1.4. Hypothesis and research Questions	6
1.5. Contribution	8
1.6. Document outline	9
2. Background	10
2.1. Semantic web.....	10
2.2. Ontology	14
2.2.1. Healthcare Ontologies.....	15
2.3. Linked data.....	16
2.4. RDF data representation.....	21
2.4.1. RDF vocabularies	22
2.4.2. SPARQL	24
2.5. Related work	25
3. Data management.....	30
3.1. Data sources	32
3.2. Data modeling, preparation and conversion.....	33
3.2.1. Vocabularies.....	34
3.2.2. URI Patterns.....	36
3.2.3. Geographical data preparation	37
3.2.4. Statistical data Preparation	38
3.2.5. Provenance Allocation	39

3.3. Data storage	40
3.4. Data License	42
3.5. Data Enrichment	42
3.6. Data Interlinking	42
4. System Overview.....	46
4.1. General overview.....	46
4.2. LOHD System Architecture	47
4.2.1. Data Layer.....	49
4.2.2. Transformation layer	50
4.2.3. Service layer	51
4.2.4. Presentation Layer	53
4.2.4.1. Visualization	53
4.3. Linked data search Engines.....	64
4.4. Case Studies of queries and visualizations.....	60
5. Conclusion and lessons Learned.....	63
5.1. Summary.....	63
5.2. Lessons learned and future works	64
6. References	65

Index of table

Table 1: Manual data interlinking results

Table 2: Automatic data interlinking results

Table 3: Sample triple data in Fuseki triple store

Index of Figure

Figure 1: Conceptual stack for semantic web

Figure 2: Linked open data cloud

Figure 3: SPARQL query interface

Figure 4: General methodology

Figure 5: home page of LOHD system

Figure 6: LOHD system architecture with layered approach

Figure 7: Query processing in the transformation layer

Figure 8: Sample SPARQL Query

Figure 9: Sample visualizations using SGVIZLER over the
LOHD System

Figure 10: sig.ma search engine over LOHD data

Figure 11: Time serous visualization over LOHD system

Figure 12: geographical visualization over LOHD system

Figure 13: Indicator based correlation visualization over LOHD
System

1. Introduction

Information is a foundation for effective decision-making. For better information, efficient management and representation of data in reusable and flexible manner is necessary. If we look around, we are surrounded by—giga bytes of data on our home computer about our day-to-day activity, dozen shelves of data at our office or data on the web, which is uploaded by different users for different purposes. With the current billion of users of social network on the web, there is plenty of data on the web. Increasing number of individuals and organizations are contributing to this open data initiative by choosing to share their data to others including healthcare organizations. There are different sources, which generate plenty of health information data starting from individuals to official sources like WHO, UNAIDS or CDC which give different health data to support the availability of health information for care givers and decision makers. Health information has been variously described as the “foundation” for better health, as the “glue” holding the health system together, and as the “oil” keeping the health system running [1]. There is however a broad consensus that a strong health information system (HIS) is an integral part of the health system, which involves the participation of different actors [2,3].

Information

In order to provide better health information for those who need it for their decision, it is evident that we need different health data representation, analysis, and querying and displaying methods. However, health information System need and service is complex. On the demand side, there are different users and uses of information – people and patients, communities, service providers, programme managers, policy-makers, providers of funds, global agencies and organizations. All need information on a range of health-measurement areas including mortality and morbidity rates; disease outbreaks; determinants of health (such as nutrition, environment, and socioeconomic status); access, coverage and quality of services; costs and expenditures; and equity sometimes with their own format and standards. On the supply side, various tools and methods are available including vital registration and census systems; household, facility and district surveys; routine clinic-based data; disease surveillance systems; national health accounts; and modeling. [4]

Health
information

Unfortunately, supply and demand in the health information field are not currently in equilibrium, with an oversupply of data coexisting with large unmet needs for information. Although many countries including sub-Saharan African countries now have relatively good data on levels of (and trends in) child mortality, health services coverage, and health determinants through donor supported programs, the available information is not modeled and visualized in a way which is useful for stakeholders of healthcare specially in developing countries. In addition, the quality of health data is often highly variable with little standardization across definitions and methodologies, and considerable overlap and duplication between it. [5]

The integration of health data across service systems is another challenge. Indeed, health data are very heterogeneous and health standards¹ have a wide variability in their implementation, and thus, we need to come up with ways that can handle this challenges. This thesis focus on how to use the World Wide Web to enhance health information data representation, query visualization and knowledge discovery as it revolutionize the way data is discovered, accessed, integrated and used on the web.

The World Wide Web², which is a system of interlinked documents, has changed the way we share information among people and systems. The World Wide Web has radically altered the way we share and access knowledge, by lowering the barrier to publishing and accessing documents as part of a global information space. This functionality has been enabled by the generic, open and extensible nature of the Web, which is also seen as a key feature in the Web's unconstrained growth for decades. [5]

Nowadays the Web has more than 100 billion documents and more than one billion people are using the Web. Information retrieval in such large-scale information system is a non-trivial task [5]. The Semantic Web is, from a pragmatic point of view, a framework of standards specified by the World Wide Web Consortium (W3C) that allows data to be shared and reused on the Web that will be discussed in detail in chapter 2.

¹ <http://www.hln.com/expertise/hit/hie/hie-standards.php>

² <http://www.w3.org/>

1.1. Motivation

Generating reliable information from a set of data and developing tools, which can support health professionals to access and understand health data properly for decision-making, has always-deep interest in me. As Hans Rosling mentioned in his “Database Hugging Disorder open data speech” at World Bank on May 2010³, we need to open data Silos and make data talk⁴. Healthcare funding, service, monitoring and prevention activities will be more effective if health professionals and decision makers get properly analyzed, interpreted and visualized information to evaluate and monitor what they have done and what they have to do.

For the last couple of decades, HIV/AIDS is an unprecedented epidemic and public health emergency in the world, especially in sub Saharan Africa. Presently, worldwide, it is estimated that over 33 million people are infected with HIV, and over 16 million have died of AIDS-related illnesses. [6] In many resource-poor countries and among marginalized groups of people in industrialized countries, the number of new HIV infections continues to rise. In some countries in Africa, AIDS related morbidity and mortality are causing major reversals in development, childhood mortality, and survival and life expectancy. Multiple National and international initiatives are taking place at least to prevent its spread and to provide ART service for those who live with the disease. [7]

In order to effectively understand the epidemic pattern, effectiveness of interventions and for better healthcare planning and decision-making, there is a need to represent HIV data and visualize it in a better and understandable way by healthcare workers, professionals, policy makers as well global funding organizations. As a student who have both health and informatics background, I am interested to do on how effectively we can represent and visualize such a data and how we can build a query system which can give intelligent search result to end users who are in need of such information for decision making. My personal feeling is “ Google” is not currently enough for health information access. We need a better representation and retrieval systems, which are intelligent for user queries.

Burdon of HIV data management and the need for better data representation tools

³ <http://blogs.worldbank.org/dmblog/videos/open-data-for-an-open-world>

⁴ <http://www.cancer.gov/cancertopics/cancerlibrary/MDT-Workbook.pdf>

The Linked Open Health Data system infrastructure will make it easy to get the information we need by linking and searching across open data sources in order to identify novel and meaningful correlations and mechanisms within the data.

Linked Open Data
as a solution

By building on Web infrastructure (URIs and HTTP), Semantic Web standards (such as the Resource Description Framework and RDF Schema [RDFS]), and vocabularies, linked data can effectively reduce barriers to data publication, consumption, and reuse, adding a rich layer of fine-grained, structured data to the Web in a layered approach. [7]

Healthcare specifically HIV/AIDS research has a wealth of available data sources to help elucidate the complex nature that lead to the global burden of the disease. However, the heterogeneous nature of these data and their widespread distribution over journal articles, proceedings, patents and numerous databases makes searching and pattern discovery⁵ a tedious and manual task.

In this thesis we will develop a system for health data representation, query in a heterogenous environment and ways to visualize and retrieve useful information for health professionals and decision makers.

1.2. Problem statement

There is huge amount of health data available on-line now than ever before either in structured or unstructured format. For example, World health organization had established a data repository⁶, which provides access to over 50 datasets on priority health topics including mortality and burden of HIV/AIDS in different WHO regions. Additionally, United Nations and CDC, have an online data repository on different indicators of different countries. Yet, whilst these initiatives are beginning to pay off - there is at the same time relatively little attention to formats and best practices. Data is published, granted, but formats vary from simple HTML tables to data in proprietary formats such as PDF and Excel, making it hard to combine,

Online official data
Challenges

⁵ <http://www.patterndiscovery.com/>

⁶ <http://apps.who.int/ghodata/>

compare and reuse information on a large scale. Additionally, even though different HIV indicators have relation between each other and other related health data, the datasets are not linked together. Vocabularies and data formats are unfamiliar and inconsistent, especially when crossing country boundaries. Finding, assembling, and normalizing these data sets is time consuming and prone to errors and, currently, no tools are implemented in WHO dataset to make intelligent queries or reasonable inferences across it.

1.3. Solution Approach

WWW allows access to a large number of valuable resources, mainly designed for human use and comprehension. The connection between documents is done by hyperlinks. With hyperlinks, search engine are able to find structure between documents. Thus, allows user to find particular structured information in a document. Human readers are capable of deducing the role of the links and are able to use the Web to carry out complex tasks. However, a computer cannot accomplish the same tasks without human supervision because Web pages are designed to be read by people, not by machines. In this age of data overload, we need methods and tools that enable us to process the data by machines.

Despite the inarguable benefits the Web provides, until recently the same principles that enabled the Web of documents to flourish have not been applied to data. We can hence understand the Web, as we know it today to be kind of incomplete.

The World Wide Web Consortium (W3C)⁷ advocates a solution based on the Linked Open Data paradigm (LOD)⁸ which is a recommended best practice for exposing, sharing, and connecting pieces of data, information, and knowledge on the Web using the RDF⁹ data model (about which more will be said in Section 2.4).

⁷ www.w3c.org

⁸ <http://linkeddata.org>

⁹ <http://www.w3.org/RDF/>

The automatic processing of information from the World Wide Web requires that data is available in a structured and machine-readable format. The Linking Open Data initiative¹⁰ actively promotes and supports the publication and interlinking of so-called Linked Open Data from various sources and domains. Its main objective is to open up data silos and to publish the contents in a semi-structured format with typed links between related data entities. As a result a growing number of Linked Open Data sources are made available which can be freely browsed and searched to find and extract useful information.

To fully benefit from Open official data, it is crucial to put information and data into context that creates new knowledge, reusability and enables powerful services and applications. This thesis use the concepts and principles of linked open data to test health data representation, querying and visualization of HIV data, of sub Saharan African countries and interconnect it with other contents which are already on the web in a way that best represent its full conceptual content and allows both automated integration and data driven decision-making. In line with this, challenges of health data representation, preparing them for query in a heterogonous environment and ways to visualize them will be assessed.

1.4. Research Questions

The primary goal of the Linked Data movement is to make the World Wide Web not only useful for sharing and interlinking documents, but also for sharing and interlinking data at very detailed levels. The movement is driven by the hypothesis that these technologies could revolutionize global data sharing, integration and analysis, just like the classic Web revolutionized information sharing and communication over the last two decades. It is contended that the deployment of Linked health Data comes with a specific set of requirements. The degree in which the requirements are accomplished (i.e. data representation, query and visualization) predetermines the system's usefulness for data consumers. Therefore, from this central aspect the following research questions arise that are dealt with in the course of this thesis.

¹⁰ Linkeddata.org/

Research Question

Is Linked open data a potential alternative for health data representation, querying and visualization?

Sub-Research Questions

- 1.1. How we can represent health data using linked open data principles?
- 1.2. Which underlying LOD technologies and tools are already available to represent, query and visualize health data?
- 1.3. How we can integrate those tools to develop a Linked Data based Health information system?

In order to answer the above research questions, the proposed approach is to try to develop a linked data based health data representation, Querying and visualization system and tests it with already available dataset. For that we choose the WHO dataset which is already available online in spreadsheet format and develop a linked data based system with the following four major Tasks.

- Represent HIV official data using RDF following linked open data principles.
- Interlink related data with already available datasets
- Develop a SPARQL interface, which is used to retrieve the data by the users.
- Develop a way to visualize a SPARQL result for better understanding by end users.

Approaches to
Answer the
research Question

1.5. Contribution

We have developed an ontology based linked open health data system, (LOHD), which allows end users to easily construct and run useful queries across multiple data sources. LOHD is flexible enough to allow queries relevant to wide range of geographical, statistical and HIV related topics. It is designed to be accessible to users who are not familiar with the structure of the underlying ontologies used in describing the datasets or with the SPARQL query language used to query the data. The system allows users to write their own query and make live visualizations using the available options of tools in the system and those who are not familiar with SPARQL queries can browse the data using linked search engines. Specifically, The major contributions of this thesis work are as follows:

Linked health data triples:

We produce more than 21,000 triples by triplication and interlinking in RDF format, which is reusable for any one interested. It will be available on the SPARQL endpoint for download.

Linked open health data system

We have developed a linked open data system, where users can query and make live visualization on the query results.

Identification of available tools and technologies

We had done Identification of required components and available technologies and tools for health data representation, Querying and visualization to create similar future LOHD projects.

Linked Data Pages

We successfully integrate our dataset with linked data search engines so that users can search- like Google- for any information they want which is already available on the dataset

Best practices

Summary of the lessons learned from development pitfalls, workarounds, and best practices on data retrieval, modeling, and integration to publishing.

Outputs

1.6. Document outline

Linked data is a new way of information system development method in which its effectiveness and usefulness is totally dependent on how much effectively we manage the data handling and publishing process. Keeping in mind this, in this thesis we discussed the different tools we use for its development and the different stages so that others can use it for the development of similar projects.

Chapter 2 presents background material for this document. It covers the essentials of semantic web, ontology and linked data, which are backbone technologies for linked open data, and summarizes its core components. The Linked Data design principles as well as semantic web, which are pillars of linked data system development are explained in order to draw a bridge. Existing Linked data systems, models and user evaluations are also summarized in order to provide the basis of the problem space. Chapter 3 contains an overview of the data management methodology from data preparation to conversion and publishing in linked data format, the core proposal of this document. Chapter 4 then outlines the overall system architecture with actual implementations of the technologies where Linked Data Pages and visualization tools are put forward as contributions.

Chapter 5 concludes by analyzing the work described, results achieved, lessons learned, and the future work up ahead and by discussing why linked data based health information representation is better than current spreadsheet representation by taking the African HIV data representation as a case study.

2. Background

2.1. Semantic web

"The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation."

Tim Berners-Lee, James Hendler, Ora Lassila (2001)

The Semantic Web is an extension of the current World Wide Web, not a separate set of new and distinct websites. It builds on the current World Wide Web constructs and topology, but adds further capabilities by defining machine-processable and understandable data and relationship standards along with richer semantic associations. Existing sites may use these constructs to describe information within web pages in ways more readily accessible by outside processes such as search engines, spider searching technology, and parsing scripts. Additionally, new data stores, including many databases, can be exposed and made available to machine processing that can do the heavy lifting to federate queries and consolidate results across multiple forms of syntax, structure, and semantics. The protocols underlying the Semantic Web are meant to be transparent to existing technologies that support the current World Wide Web. [7]

Information on the Web is becoming increasingly fragmented and varied in terms of appropriateness, timeliness, accessibility, and trustworthiness. Search engines are wonderful tools but, increasingly, fault lines are appearing especially in health information needs. These fault lines manifest themselves in doubts about completeness of search; the growing use of script-like search commands such as "filetype"; or the rise in search engines focusing on specific types of data needs for example, how much is the prevalence of HIV in West Africa where the user need specific answers. [8]

Challenges in
information search

Two new data and logic structures recently approved by the World Wide Web consortium (W3C)¹¹ are making it possible to make information richer and more autonomous and, ultimately, far more accessible and adaptive. These new constructs as explained by Semantic Interoperability working group [8], are – Resource Description Framework (RDF) and Web Ontology Language (OWL) – make extensive use of knowledge representation principles to add additional functionality and compatibility to existing W3C markup languages. RDF provides a framework for establishing relationships between data, whereas OWL enhances RDF with the ability to specify constraints on different data elements and their relationships to one another. These standards – in conjunction with new tools and infrastructure components built to support them – are driving the development of adaptive computing within the enterprise as well as the growth of the next generation of the web, called the Semantic Web. [8]

OWL and RDF

The vision of the Semantic Web is to extend the current web by enriching the information transmitted and accessed over the Internet with well-defined meaning, thus enabling computers to do more of the work in assembling and processing data in order to turn it into highly relevant information and knowledge. In other words, the initiatives underlying the Semantic Web establish a set of protocols and technologies that promise to improve the categorization and association of data thereby enhancing the ability to create relationships and to generate inferences among diverse systems and data. [9]

Vision of the
semantic web

According to the World Wide Web Consortium (W3C), the Web can reach its full potential only if it becomes a place where data can be shared, processed, and understood by automated tools as well as by people.

¹¹ www.w3c.org

Semantic web address problem of information overload, stovepipe systems, and poor content aggregation [11]. The fundamental roots of these problems are the lack of semantic definitions in individual systems, the lack of semantic integration among data sets, and the lack of semantic interoperability across disparate systems. The Semantic Web extends beyond the capabilities of the current Web and existing information technologies, enabling more effective collaborations and smarter decision-making. It is an aggregation of intelligent websites and data stores accessible by an array of semantic technologies, conceptual frameworks, and well-understood contracts of interaction to allow machines to do more of the work to respond to service requests – whether that be taking on rote search processes, providing better information relevance and confidence, or performing intelligent reasoning or brokering. [7]

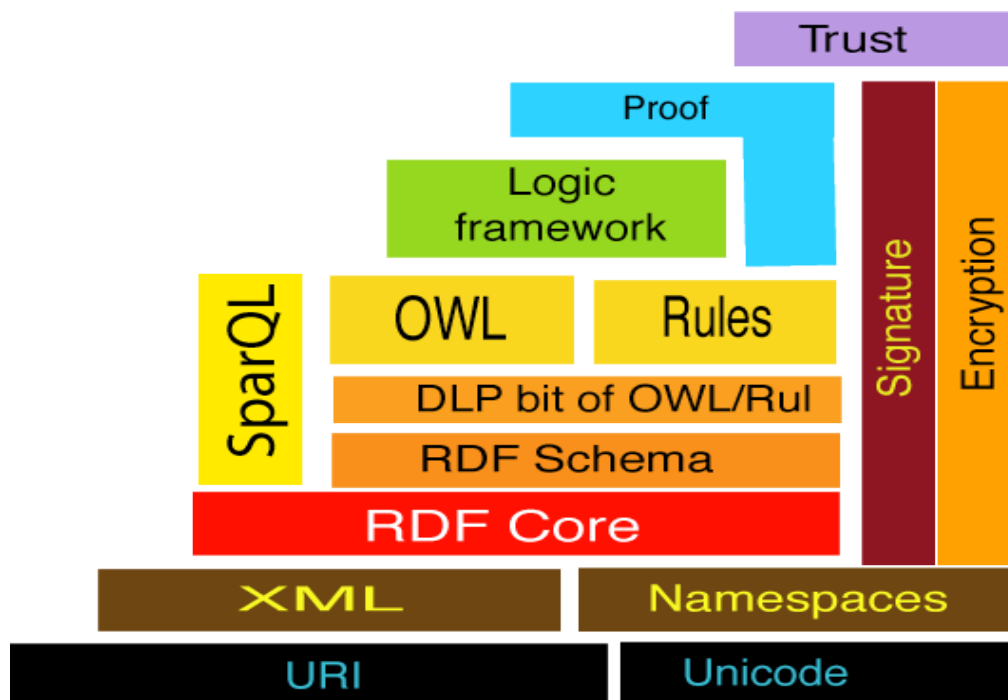


Figure 1: A conceptual stack for the Semantic Web ¹²

¹²Adapted from <http://www.w3.org/2000/Talks/1206-xml2k-tbl/>

The application of semantic technology to the medical domain will provide IT systems the ability to better understand terms and concepts as data is transmitted from one system to another, while preserving the meaning of the content. For this process to work effectively, different initiatives are being made involving the classification of medical terms and their meanings in medical experts initiatives. Tools in this area make use of classification systems that produce controlled vocabularies, lexicons, taxonomies and ontologies. [8]

Semantics in
healthcare

For humans, the meaning of a given word is normally obtained by consulting a dictionary or by looking at the context where the word is being used. The computer does not make use of textual dictionary definitions and has no pre-existing repository of contexts, but instead requires a semantic representation that is simpler and more precise. Natural language processing systems represent the meaning of a given word or phrase using a symbol or code. For an Electronic Medical Record (EMR)¹³ system, “heart” and “cardiac” are two unrelated terms. For humans, however, both terms have the same 'semantic' meaning. To create such understanding between systems we use vocabularies, which insure interoperability in communication. [10]

Semantics
challenge on the
web

One of the most important ways semantic interoperability services and resources in healthcare can be used relates to reconciling clinical data contained in diverse EHR systems. Semantic interoperability is a concept that will definitely contribute to improvement in health care over time because it will deliver the right meaning of medical terminology to each collaborating system user every time, via a service-oriented web-based solution. [11]

¹³ *Semantics is a big challenge in electronic medical record systems especially in health information retrieval. Now days there are different organizations that are doing standardization like HL7 AND SNOMED.*

2.2. Ontology

Ontology¹⁴ is a data model that represents a set of concepts within a domain and the relationships between those concepts explicitly [9]. It is used to reason about the objects within that domain. Ontologies are used in artificial intelligence, the semantic web, software engineering, biomedical informatics, and information architecture as a form of knowledge representation about the world or some part of it. [10]

Ontology defines the terms and concepts (meaning) used to describe and represent an area of knowledge. It may be conceived as an explicit specification of a conceptualization that includes a set of objects, their properties and their values along with the describable relationships among them, reflected in a representational vocabulary with which a knowledge-based program represents knowledge. [11]

One use of ontologies is to externalize a model and make it easier for business applications to share or reuse knowledge and improve information navigation and search by using reasoning, which is one of the important concepts in healthcare system development. Furthermore, the externalization of models facilitates customization of an application without modifying code. In ontologies, definitions associate the names of entities in the universe of discourse (e.g., classes, relations, functions, or other objects) with human-readable text describing what the names mean, and formal axioms that constrain the interpretation and well-formed use of these terms. [10]

Uses of ontology

¹⁴ <http://www-ksl.stanford.edu/kst/what-is-an-ontology.html> Tom Gruber give a widely used definition, as “An ontology is a specification of a conceptualization”. That is, ontology is a description (like a formal specification of a program) of the concepts and relationships that can exist for an agent or a community of agents.

2.2.1. Healthcare Ontologies

Different types of health related ontologies have been developed with different initiatives to insure interoperable healthcare communications. A healthcare semantic web initiative white paper assesses the different on-going initiatives in this aspect. [10]

The *Disease Ontology*¹⁵ is implemented as a directed acyclic graph (DAG) and utilizes the Unified Medical Language System (UMLS) as its immediate source vocabulary to access medical Ontologies such as ICDCM. Using this standard, much of the process of updating the ontology can be handled by UMLS vocabulary, freeing resources for clinicians to pursue more urgent tasks. [10]

Healthcare
ontology initiatives

The OASIS Health Monitoring Ontology can make use of some of the concepts and structures of those Ontologies, particularly applicable in the Health Monitoring and surveillance field. In addition, the OASIS Health Monitoring Ontology can describe the Health Monitoring domain by means of appropriate terms derived from ontology-like vocabularies such as “NCI Thesaurus”, “UMLS Knowledge Sources” and “SNOMED”. [10]

The *Unified Medical Language System*¹⁶ (UMLS) facilitates the development of computer systems that behave as if they "understand" the language of biomedicine and health. Developers use the Knowledge Sources (databases) and tools to build or enhance systems that create, process, retrieve, and integrate biomedical and health data and information. The Knowledge Sources are multi-purpose and are used in systems that perform diverse functions involving information types such as patient records, scientific literature, guidelines, and public health data. The associated software tools assist developers in customizing or using the UMLS Knowledge Sources for particular purposes. The lexical tools work more effectively in combination with the UMLS Knowledge Sources, but can also be used independently. [10]

¹⁵ <http://diseaseontology.sourceforge.net/>

¹⁶ <http://www.nlm.nih.gov/pubs/factsheets/umls.html>

*SNOMED Clinical Terms*¹⁷ (SNOMED CT) is a dynamic, scientifically validated clinical health care terminology and infrastructure that makes health care knowledge more usable and accessible. It provides a common language that enables a consistent way of capturing, sharing and aggregating health data across specialties and sites of care [13].

Ontologies can help in building silo-busting applications that need to link data items (datum) to other data items (as opposed to web page to web page) over the web in order to perform entity correlation (or entity resolution). A datum can be a row in a relational database and technologies exist to provide an RDF view over a relational database table. The RDF view itself can be defined in terms of an OWL ontology or RDFS vocabulary. Hence, LOD can integrate data across health applications and organizations by providing a semantic query and visualization layer on top of existing applications. [11]

2.3. Linked data

Linked Data is a data publishing technique that uses common Web technologies to connect related data and make them accessible on the Web. It relies mainly on identifying resources with (HTTP) Uniform Resource Identifiers (URI), and, using standards such as the Resource Description Framework (RDF), by providing data about these resources and connecting them to other resources on the Web. [5]

When data were structured and organized as a collection of records in dedicated, self-sufficient databases, information was retrieved by performing queries on the database using a specialized query language; for example SQL (Structured Query Language) for relational databases or OQL (Object Query Language) for object databases. In modern healthcare, exploiting the different kinds of available information about a given topic is challenging because data are spread over the World Wide Web (Web),

¹⁷ <http://www.ihtsdo.org/snomed-ct/>

hosted in a large number of independent, heterogeneous and highly focused resources. [5]

Hands-off data handling requires moving from a Web of documents, only understandable by humans, to a Web of data in which information is expressed not only in natural language, but also in a format that can be read and used by software agents, thus permitting them to find, share and integrate information more easily [5]. In parallel with the Web of data, which is focused primarily on data interoperability, considerable international efforts are ongoing to develop programmatic interoperability on the Web with the aim of enabling a Web of programs. Here, semantic descriptions are applied to processes, for example represented as Web Services. [9]

Interoperability

The Linked Data efforts are concerned with publishing and querying all sorts of data that is interconnected in the form of Tim Berners-Lee's Giant Global Graph. Some of the motivations behind this are to uncover insights about societies, build smarter systems, making predictions, democratizing data for people, or to make better decisions. [5]

Linked Data is a new field of research to be used as a representation method for complex data. The term Linked Data is used to refer to a set of best practices for publishing and connecting structured data on the Web. [5] It explicitly encourages the use of dereferenceable links using Linked Data principles. Linked Data is implemented by various type of technology such as RDF and XML. There are search engines that allow crawling through this web of data and perform query results from user. Many domains also make use of this technology. Now there are millions of existing ontological vocabularies and datasets for geographical data, and space and time-related data that can insure interoperability and richness of the data. [13] All those data, which are developed using those standardized vocabularies, are uploaded on the linked data cloud. The cloud is being expanded from year to year by adding data from different sectors by different initiatives.

representation for requested resources which make the use of http not only for connecting documents but also to connect data over the web.

When organizations, either healthcare or governmental organizations publish their data in general and in Linked Open Data format in particular, Tim Berners-Lee suggested a 5-star deployment scheme in 2009¹⁹ [which, slightly modified, looks like this:

- * Make data available on the Web (whatever format) under an open license
- ** Make it available as structured data (e.g. Excel instead of image scans)
- *** Use non-proprietary formats (e.g. CSV instead of Excel)
- **** Use URIs to identify data items, so that they can be referenced on the Web
- ***** Link your data to other's data to provide context

Tim Berners-Lee's
5 star linked data

The Linked Data design pattern is based on an open world assumption, uses dereferenceable HTTP URIs for identifying and accessing data items, RDF for describing metadata about those items, and semantic links to describe the relationships between those items. Other standards used in LOD applications include RDFS (for describing RDF vocabularies) and SPARQL (for querying RDF graphs). Using those technologies as a common platform, and with better data representation methods, it is possible to create a big data cloud. As mentioned earlier, the linked data cloud is increasing from year to year. Among the available data, there are a lot of initiatives in healthcare, but there is no data cloud exclusively to publish HIV data on the web. In the following section, we will discuss how RDF can represent health information data flexibly so that the data can be reusable and available on the cloud. [9]

¹⁹ <http://www.w3.org/DesignIssues/LinkedData.html>

Additionally, as explained by different projects and individuals representing data in RDF and publishing them in linked open data, have the following benefits [12,13,14,15,16]

Federation:

All datasets published as Linked Data share a uniform model, the RDF statement data model. With this data model all information is represented in facts expressed as triples consisting of a subject, predicate and object. The components used in subject, predicate or object positions are mainly globally unique IRI/URI entity identifiers. At the object position also literals, i.e. typed data values can be used. Additionally the predicates are mostly expressive of the data type, which make understanding relations really easy.

De-referencability:

URIs are not just used for identifying entities, but since they can be used in the same way as URLs they also enable locating and retrieving resources describing and representing these entities on the Web.

Coherence:

When an RDF triple contains URIs from different namespaces in subject and object position, this triple basically establishes a link between the entity identified by the subject (and described in the source dataset using namespace A) with the entity identified by the object (described in the target dataset using namespace B). Through the typed RDF links data items are effectively interlinked.

Integrability:

Since all Linked Data source share the RDF data model, which is based on a single mechanism for representing information, it is very easy to attain a syntactic and simple semantic integration of different Linked Data sets. A higher-level semantic integration can be achieved by employing schema and instance matching techniques and expressing found matches again as alignments of RDF vocabularies and ontologies in terms of additional triple facts.

Timeliness:

Publishing and updating Linked Data is relatively simple thus facilitating a timely availability. In addition, once a Linked Data source is updated it is straightforward to access and use the updated data source, since time consuming and error prone extraction transformation and Loading is not required. However, these advantages of pursuing a Linked Data integration approach cannot be realized immediately, but will be attained by small iterative integration and refinement.

2.4. RDF data representation

The Resource Description Framework (RDF)²⁰ is a general-purpose language for representing information in the Web. It is particularly intended for representing metadata about Web resources, but it can also be used to represent information about objects that can be identified on the Web, even when they cannot be directly retrieved from the Web. To some extent, RDF is a lightweight ontology language to support interoperability between applications that exchange machine-understandable information on the Web. RDF has a very simple and flexible data model, based on the central concept of the RDF statement. We also consider the concept of vocabulary as part of the RDF data model, due to its relevance to ontology modeling. RDF offers three equivalent notations: RDF triples, RDF graphs, and RDF/XML. The RDF triples notation translates RDF statements directly into character strings. [11]

The key ingredient in the information that is returned to the user has to do with the model of the data in the response. Regardless of the syntax that is used, Resource Description Framework (RDF) is essentially an entity-relationship model that provides a way to make statements about the things in our reality. A statement contains three atomic parts, also known as a triple: the subject resource which the statement is about, followed with a

²⁰ <http://www.w3.org/RDF/>

property which is a vocabulary term that describes the type of relationship it has to an object resource. [5] Each of these components is typically represented using HTTP URIs, with the possibility of the object resource being a literal string. In mathematical terms, RDF is a directed, labeled graph, which conceptually depicts a graph of things. What makes this method to make claims about things worthwhile is the act of linking any two resources identified through URIs together in a particular way. It fundamentally presents an opportunity to discover new resources in uniform way, whether the resource is in local storage or somewhere else. [12]

2.4.1. RDF vocabularies

A RDF statement, also called a triple in RDF terminology is an association of the form (*subject, predicate, object*). A *triple* is also known as a “statement” and is the basic “fact” or asserted unit of knowledge in RDF. Multiple statements get combined together by matching the *subjects* or *objects* as “nodes” to one another (the *predicates* act as connectors or “edges”). As these node-edge-node triple statements get aggregated, a network structure emerges, known as the RDF *graph*. The referenced “resources” in RDF triples have unique identifiers, IRIs, that is Web-compatible and Web-scalable. These identifiers can point to precise definitions of predicates or refer to specific concepts or objects, leading to less ambiguity and clearer meaning or semantics. The subject of a RDF statement is a resource identified by a Uniform Resource Identifier (URI). [5]

Triple format

In RDF triple statements, properties are vocabulary terms that are used to relate a subject to an object. As these resources are accessible via HTTP URIs, when dereferenced they provide a description for the term in use. Some of the well-known vocabularies that are used in Linked Data publishing include: Friend of a Friend (FOAF)²¹; to describe people and the things that they do, RDF Data Cube vocabulary²²; which is used to describe

²¹ <http://xmlns.com/foaf/spec>

²² <http://www.w3.org/TR/vocab-data-cube>

multi-dimensional statistical data, Simple Knowledge Organization System (SKOS)²³ to describe controlled thesauri, classification schemes and taxonomies, DCMI Metadata Terms (DC Terms)²⁴ for general purpose metadata relations, and¹² Vocabulary of Interlinked Datasets (VoID)²⁵; to provide metadata on target datasets.[11]

Common RDF
Vocabularies

RDF Schema (RDFS)²⁶ and the Web Ontology Language (OWL)²⁷ are used to explicitly represent the meanings of the resources described on the Web and how they are related. These specifications, called ontologies, describe the semantics of classes and properties used in Web documents. Ontology suitable for the example above might define the concept of disease (including its relationships with other concepts) and the meaning of the predicate “*is located on*”. [12] In an ideal world, each ontology should be linked to a general (or top-level) ontology in order to enable knowledge sharing and reuse. In the domain of the Semantic Web, several ontologies have been developed to describe Web Services and applications.[9]

RDF triples can be applied equally to all structured, semi-structured and unstructured content. By defining new types and predicates, it is possible to create more expressive vocabularies within RDF. This expressiveness enables RDF to define controlled vocabularies with exact semantics. These features make RDF a powerful data model and language for data federation and interoperability across disparate datasets [12] especially in a heterogeneous healthcare data.

Triple format

²³ <http://www.w3.org/2004/02/skos/>

²⁴ <http://dublincore.org/documents/dcmi-terms/>

²⁵ <http://www.w3.org/TR/void/>

²⁶ <http://www.w3.org/TR/rdf-schema/>

²⁷ <http://www.w3.org/TR/owl-features/>

2.4.2. SPARQL

SPARQL Protocol and RDF Query Language (SPARQL)²² are a protocol and a query language to retrieve and manipulate RDF data. It can be used to express queries across local and remote data sources, whether the data resides in RDF files or databases. SPARQL queries consist of graph patterns written in a fashion similar to Turtle (an RDF format), and allow modifiers for the patterns. In the Linked Data community, it is common to see publicly accessible SPARQL endpoints where queries are sent and received over HTTP. [12]

SPARQL²⁸ is a query language for RDF. A SPARQL query is represented by a graph pattern to match against the RDF graph. Graph patterns contain triple patterns that are like RDF triples, but with the option of query variables in place of RDF terms in the subject, predicate or object positions. SPARQL saves development time and cost by allowing client applications to work with only the data they're interested in. [13]

Example: Find countries population, GDP, and HIV prevalence, in order to determine if there is a relationship between population density and HIV prevalence. Without SPARQL, you might tackle this by writing a first query to pull information from population' pages on Wikipedia, a second query to retrieve GDP data from another source, and then another source to find the prevalence of HIV. With SPARQL, this application can be accomplished by writing a single SPARQL query that federates the appropriate data source. The application developer need only write a single query and no additional code for each repository.

SPARQL example

SPARQL builds on other standards including RDF, XML, HTTP, and WSDL. This allows reuse of existing software tooling and promotes good interoperability with other software systems. Examples: SPARQL results are expressed in XML: XSLT can be used to generate friendly query result displays for the Web It's easy to issue SPARQL queries, given the abundance of HTTP library support in Perl, Python, php, Ruby, etc. [15]

²⁸ <http://www.w3.org/TR/rdf-sparql-query/>

All the above were the common underlying technologies in linked open data systems. There are different individual and institutional initiatives in building linked data systems on the web. In the following section, we will revise those initiatives, which have a direct contribution for this thesis work. Then in the third chapter we will see how we manage our data in our linked health data system and the steps from data preparation to data publishing and visualization.

2.5. Related work

Health care is an information-intensive science and research in health care service; monitoring and evaluations of global funding for priority diseases and their effectiveness heavily depend on the availability and the efficient use of information. To make the information available in robust RDF representation, there are already a number of efforts to convert health care and life science related data sets to Linked Data such as LODD, LinkedCT, OBO ontologies and the W3C's Health Care and Life Sciences working groups.

The Semantic Web Health Care and Life Sciences (HCLS Interest Group)²⁹ is established by the World Wide Web Consortium (W3C) to support the use of Semantic Web technologies in health care, life sciences, clinical research and translational medicines. The group focuses on aiding decision-making in clinical research, applying the strengths of Semantic Web technologies to unify the collection of data for the purpose of both primary care (electronic medical records) and clinical research (patient recruitment, study management, outcomes-based longitudinal analysis, etc.). Subgroups, on the other hand, focus on making the biomedical data available in RDF; dealing with biomedical ontologies, focus on drug safety and efficacy communication and support researchers in the navigation and annotation of the large amount of potentially relevant literature. [13]

²⁹ <http://www.w3.org/blog/hcls/>

U.S. National Library of Medicine – NLM³⁰ under National Institutes of Health as the world's largest medical library provides information (i.e. Medical Subject Headings – MeSH⁷) and research services like National Center for Biotechnology Information – NCBI³¹. NCBI provides access to biomedical and health information through semantic resources like PubMed³² Online Mendelian Inheritance in Man – OMIM³³ and many others. Based on semantically structured and uniquely identified library sources a semantic data integration platform **LinkedLifeData**³⁴ for the biomedical domain is developed. [14]

LODD, i.e. the Linking Open Drug Data project³⁵ mainly converts, publishes and interlinks drug data that is available on the web, ranging from impacts of drugs on gene expression to results of the clinical trials. A number of datasets have been converted in this project including DrugBank, DailyMed, SIDER to name a few. However, these datasets are restricted to drug data and even though they do contain disease data (from the DisEasome dataset), they do not connect the number of deaths or the health expenditure or the status of the health system in each country for each of the diseases that are included. [15]

DO - Disease Ontology³⁶ is another (open source) project, semantically integrates these medical and disease vocabularies through extensive cross mapping between DO terms as well. [14]

³⁰ [<http://www.nlm.nih.gov/>]

³¹ <http://www.ncbi.nlm.nih.gov/>

³² <http://www.ncbi.nlm.nih.gov/pubmed/>,

³³ [<http://www.ncbi.nlm.nih.gov/omim>]

³⁴ <http://linkedlifedata.com/>

³⁵ <http://www.w3.org/wiki/HCLSIG/LODD/>

³⁶ <http://www.disease-ontology.org/>

Disease³⁷ 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. [13]

LinkedCT³⁸ is the Linked Data version of Clinical- trails.gov, which publishes data about clinical trials in RDF and links it to other datasets such as PubMed. Even though, in Linked CT each trial is associated with a disease and a drug, it does not provide information about the prevalence of the disease in a particular country, which is provided in GHO³⁹. [15]

OBO is the Open Biological and Biomedical Ontologies project⁴⁰ which aims to create a suite of interoperable reference ontologies in the biomedical domain. It brings together biology researchers and ontology developers who work together to develop a set of ontologies as well as design principles that can help develop interoperable ontologies. However, most of the ontologies developed are at the experimental level or organismal level and are not yet sufficiently interlinked with other datasets available as Linked Data. [15]

Centers for Medicare and Medicaid Services⁴¹ -publishing Clinical Quality Linked Data on Health.data.gov, beginning with Hospital Compare⁴². Hospital Compare Linked Data provides reports and survey results about how well hospitals treat various conditions, each with specific metrics that apply to measures designed to give citizens an understanding of how well hospitals perform when compared with state and national statistics. What's

³⁷ <http://diseasome.eu/>

³⁸ <http://linkedct.org/about/>

³⁹ <http://www.who.int/gho/en/>

⁴⁰ <http://obofoundry.org/>

⁴¹ <http://www.cms.gov/>

⁴² <http://www.data.gov/health/blog/clinical-quality-linked-data-healthdatagov>

different about this Linked Data implementation is that the definition of each class of thing in the Hospital Compare domain (including but not limited to Hospital, Condition, Measure and Metric) and the identity of every instance of each class has a globally unique address on the world wide network of computers, independent from the temporal datasets that contain periodically sampled statistical values about them. This makes it easier to accumulate more samples about how well that specific hospital is doing over time, as subsequent publications will automatically aggregate new data around each and every domain concept and their instances.

GeoNames⁴³ is a frequently used data-hub that has geographical components published as linked open data. LinkedGeoData⁴⁴ is a wider scale effort to add spatial dimension to the web of data. It uses the information collected by the OpenStreetMap⁴⁵ project and makes it available as a knowledge base according to the Linked Data principles.[13]

Additionally, There are some individual and institutional initiatives in publishing health data using linked open data principles. Nurefsan Guer [13] develops an infrastructure by integrating geo web and semantic web which aimed on reducing the boundaries between semantic web and geo web. Use case data was taken from Valencia CSISP- Research Center of Public Health in which the mortality rates for particular diseases are represented spatio-temporally which was divided into three conceptual domains -health, spatial, statistical, enhanced with semantic relations and descriptions by following Linked Data Principles.

⁴³ *GeoNames is a geographical database that is available for download free of charge under a creative commons attribution license. <http://www.geonames.org>*

⁴⁴ *<http://linkedgeodata.org/>*

⁴⁵ *<http://www.openstreetmap.de/>*

Stats2RDF project is carried by AKSW⁴⁶– Agile Knowledge Engineering and Semantic Web research group. The whole topic of the project is stated as “Representing multi-dimensional statistical data as RDF using RDF Data Cube Vocabulary”. World Health Organization – WHO’s Global Health Observatory ⁴⁷dataset is used as an initial use case. In this work, they converted all WHO data into triple and make them available online. [15] Additionally the research group had done REDD-Observatory [14], an approach for evaluating the Research-Disease Disparity based on the interlinking and integrating of various biomedical data sources. Specifically, they develop a method for representing statistical information as Linked Data and adopt interlinking algorithms for integrating relevant datasets. The assessment of the disparity is then performed with a number of parameterized SPARQL queries by taking data from different sources including GHO. [14,15]

There are some works in link discovery using SILK ⁴⁸where different types of semantic links should be discovered between Web data sources that often mix terms from different and inconsistent vocabularies. Those includes Raimond et al [16] who propose a link discovery algorithm that takes into account both the similarities of web resources and of their neighbors. The algorithm is implemented within the GNAT tool and has been evaluated for interlinking music-related data sets. Hassanzadeh et al. [17] describe a framework for the discovery of semantic links over relational data, which also introduces a declarative language for specifying link conditions. Therefore mentioned sources above carry importance for the work and considered to be linked in further chapters.

⁴⁶ <http://aksw.org/>

⁴⁷ <http://apps.who.int/ghodata/>

⁴⁸ *The Silk framework is a tool for discovering relationships between data items within different Linked Data sources. Data publishers can use Silk to set RDF links from their data sources to other data sources on the Web.*

3.Data management

The Semantic Web main aim is to build a common framework that allows data to be shared and reused across applications, enterprises, and community boundaries. To create such a data cloud, efficient management of data is the backbone. To realize this, it proposes to use RDF as a flexible data model and use ontology to represent data semantics. Currently, relational models and XML tree models⁴⁹ are widely used to represent structured and semi-structured data. But they offer limited means to capture the semantics of data. RDFS and OWL ontologies (as explained in section 2.4.1) can effectively capture data semantics and enable semantic query and matching, as well as efficient data integration. [11] In this thesis, we will represent contents of health information using linked data principles in the form of RDF – by taking out from the deep web- and we will try to make the data usable for health information managers and decision makers.

The term Deep Web refers to Web content that cannot be directly indexed by search engines. Studies showed that Deep Web content is particularly important. However, obtaining such content is challenging and has been acknowledged as a significant gap in the coverage of search engines. [18]

Indeed, much of the information hidden in Web sites is dynamically generated from relational databases or spreadsheets, and search engines are not able to find them. Estimates suggest that the volume of data stored in data silos greatly exceeds that of the Surface Web – with nearly 92,000 terabytes of data on the Deep Web versus only 167 terabytes on the Surface Web, as of 2008. [19]

Data management
challenges

⁴⁹ <http://www.w3.org/XML/Datamodel.html>

In particular, a large part of scientific and business data today is captured using Excel spreadsheets, and remains inaccessible to search engines. [20] Typically, one provides a wrapper that allows users to query datasets and their attributes, e.g., geographical location (states, regions), temporal data (month, quarters). The problem with this approach is that the data is only accessible to the few users that are aware of it, but hidden to users at large. So linked data is stressed the necessity to surface Deep Web data, i.e., to make it visible to search engines and, thus, largely findable. [21]

The Linked Data approach, based on the Semantic Web stack of standards⁵⁰ and technologies, provides a framework in which to publish, query and consume data in the Web [22]. The RDF standard, in particular, is very useful for data surfacing, for it provides a “lingua franca ” in which heterogeneous datasets can interoperate. Unfortunately, tools that support the transformation of data stored in the Deep Web to RDF formats are still in their infancy [23]. Most of the tools focuses on the transformation of relational data, as opposed to data stored in spreadsheets, and requires the installation of additional tools and add-ons [24,25,26].

In this thesis, we show the methods and tools we use in the transformation of data stored in spreadsheets to RDF. As explained in the first section 2.4-It promotes the reuse of standard RDF vocabularies, to secure interoperability with data published in the Linked Data format. Additionally, we developed a layer based health information data representation and visualization tool that will make data understanding easy.

⁵⁰ <http://www.w3.org/standards/semanticweb/>

3.1. Data sources

There are different international organizations as well country specific organizations, which publish HIV data about countries. The dataset created for this paper is based on WHO ⁵¹ data set in its global health observatory data repository¹, and missing data for some years was complemented from other similar official sources of UNAIDS⁵² and country specific official sources. In the databases, HIV statistical data as well as additional location and total population information were extracted for sub-Saharan African countries.

Additional GDP data was extracted from World Bank ⁵³ and population data was extracted from Wikipedia ⁵⁴. Those additional data sets help as in a cross-domain query and analysis (i.e. from different sources in a different way of data representation) and to be able to use those data for analysis. For example as shown in the visualization section, we take data about ART coverage from our endpoint and a data from world bank endpoint about GDP and we were able to analyse the relationship between GDP and ART coverage rate between countries and the effect of GDP on ART coverage. Additionally silk interlinking algorithm was used to link our data with other related health domain datasets that are already on linked data cloud data hub⁵⁵.

⁵¹ <http://www.who.int/gho/database/en/>

⁵² <http://www.unaids.org/en/dataanalysis/> UNAIDS is the Joint United Nations Programme on HIV/AIDS which is an innovative partnership that leads and inspires the world in achieving universal access to HIV prevention, treatment, care and support.

⁵³ <http://data.worldbank.org/>

⁵⁴ http://en.wikipedia.org/wiki/List_of_countries_by_population

⁵⁵ <http://datahub.io/dataset/archiveshub-linkeddata>

3.2. Data modeling, preparation and conversion

In this section we go over several areas, which are at the heart of representing health data as Linked Data. Which vocabularies are reused and created, URI design patterns, and Data Cube's data structure definitions of our data for those who intend to reuse it.

In order to employ the Web as a medium for data and information integration, comprehensive datasets and vocabularies are required as they enable the disambiguation and alignment of other data and information.

A better way to secure interoperability is by a priori design, i.e., by selecting appropriate standards, if one exists, to guide the design of the data source. [11] The same philosophy is applicable to Linked Data, as stated by Bizer, Cyganiak and Heath [23] "In order to make it as easy as possible for client applications to process your data, you should reuse terms from well-known vocabularies wherever possible. You should only define new terms yourself if you cannot find required terms in existing vocabularies". Unfortunately, that is not what happens in practice. Most users prefer creating new vocabularies (as do the vast majority of triplification tools) to spending the required time and effort to search for adequate matches [21]. There are not withstanding numerous standards that designers cannot ignore when specifying triple sets and publishing their content.

Our background in databases, particularly past experiences with the construction of schema design and ontology matching, convinced us that the use of standards in schema design is the only viable way to guarantee future interoperability. [11, 22] We were anchored on this principle and strive to promote the reuse of standards by implementing a guided, four-step process starting from the data collection until the provenance allocation. The first step consists in choosing the vocabulary to be used- i.e existing vocabularies, the second step is selecting the data to be converted, the third is assigning semantics to the data and the fourth and final step consists in assigning provenance to the data generated which will make the data reuse easy.

Data modeling
challenges

3.2.1. Vocabularies

All data models use common vocabularies and the following are the common vocabularies we use in creating the linked data:

1. [RDFS](#)⁵⁶: allows you to express the relationships between things by standardizing on a flexible, triple-based format and then providing a vocabulary ("keywords" such as `rdf:type` or `rdfs:subClassOf`) which can be used to say things. This is one of the vocabularies we use to represent health data.
2. [OWL](#)⁵⁷: The OWL Web Ontology Language is designed for use by applications that need to process the content of information instead of just presenting information to humans. OWL facilitates greater machine interpretability of Web content than that supported by XML, RDF, and RDF Schema (RDF-S) by providing additional vocabulary along with a formal semantics. It shows you how to work efficiently with database queries and automatic reasoners, and it provides useful annotations for bringing your data models into the real world.
3. [FOAF](#)⁵⁸: is a project devoted to linking people and information using the Web. Regardless of whether information is in people's heads, in physical or digital documents, or in the form of factual data, it can be linked. FOAF integrates three kinds of network: social networks of human collaboration, friendship and association; representational networks that describe a simplified view of a cartoon universe in factual terms, and information networks that use Web-based linking to share independently published descriptions of this interconnected world. FOAF helps users can retain some control over their information in a non-proprietary format.

⁵⁶ <http://www.w3.org/TR/rdf-schema/>

⁵⁷ <http://www.w3.org/TR/owl-features/>

⁵⁸ <http://xmlns.com/foaf/spec/>

4. [RDF Data Cube](http://www.w3.org/TR/vocab-data-cube/)⁵⁹ is used to describe multi-dimensional statistical data. It provides a means to do this using the W3C RDF (Resource Description Framework) standard. The model underpinning the Data Cube vocabulary is compatible with the cube model that underlies SDMX (Statistical Data and Metadata exchange), an ISO standard for exchanging and sharing statistical data and metadata among organizations. The Data Cube vocabulary is a core foundation, which supports extension vocabularies to enable publication of other aspects of statistical data flows. SDMX for the statistical information model, SKOS to describe the concepts in the observations, and DC Terms for general-purpose metadata relations.

All those vocabularies were used to describe the data by using expressive and generic predicates. Some of the predicates were replaced with more generic elements from the Data Cube Vocabulary (e.g., qb:Observation instead of sdmx:Observation) or SKOS (e.g., skos:ConceptScheme in place of sdmx:code list) to make them more understandable by the potential users of the systems (health professionals and healthcare managers to be able to understand the relations more easily). We assume that using some of the terms that are already known by health professionals will make the system usable and of course adaptable.

⁵⁹ <http://www.w3.org/TR/vocab-data-cube/>

3.2.2. URI Patterns

On the Semantic Web, URIs identify not just Web documents, but also real-world objects like people and cars, and even abstract ideas and non-existing things like a mythical unicorn. We call these real-world objects or things.⁶⁰

Given such a URI, how can we find out what it identifies? We need some way to answer this question, because otherwise it will be hard to achieve interoperability between independent information systems. We could imagine a service where we can look up a description of the identified resource, similar to today's search engines. But such a single point of failure is against the Web's decentralized nature. [5]

Instead, we should use the Web itself—an extremely robust and scalable information publishing system—as a lookup service for resource descriptions. Whenever a URI is mentioned, we can look it up to retrieve a description containing relevant information and links to related data. This is so important that we make it our number one requirement for *cool* URIs:

URI patterns

A number of URI design patterns are established in recent years with similar considerations and guidance for developing and maintaining URIs. W3c Interest group⁶¹ developed a document, which explains how we can choose cool URI's on the semantic web. There are two different strategies to make URI that identify real-world objects dereferenceable. The strategies are called 303 URI and hash URI. Hash URIs are used when we want to combine the descriptions of multiple resources into one document and still keep them dereferenceable. Slash URIs should be used in combination with 303 redirects, and are good when you want to keep a one-to-one mapping between the dereferenceable descriptions of resources and documents. [5]

⁶⁰ <http://www.w3.org/TR/cooluris/>

⁶¹ <http://www.w3.org/TR/cooluris/#distinguishing>

We use of slash URIs throughout the schema and data for its observations. The reason for this is to keep the URI patterns consistent and to make sure that all important resources when dereferenced returned information. Since the content size of the responses for statistical data may be heavy, the slash URIs approach appeared to be preferable to hash URIs, as the latter would not allow distinct requests in majority of the deployments on the Web. This is independent of accessing these resources via SPARQL endpoints.

To avoid duplication of URI, we use LSRN (life science record name)⁶², which serves as a permanent home for the LSRN specification, a portal for editors and contributors, and as a LSRN-to-URN resolution service. An LSRN-based permanent URL is dual-purpose: it can be used as a regular URL for the associated record; also, it can be used as a globally unique URI for the record in Semantic Web applications. [5,25,26]

3.2.3. Geographical data preparation

The geographic component of an epidemic is an important one, as geography not only influences the spread of the disease, but also its treatment. Geography is an important factor in early attempts to understand an epidemic. [27,28] In the global response to the HIV epidemic, GIS provides an important tool in addressing such issues as areas of high transmission, [29] most-at-risk populations, [30-31] access to services, and understanding the epidemiology of the disease. [32,33,34]

GeoNames⁴⁵ 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 was done to the capital city of each country. RDF links connected with the same predicate owl:sameAs used for linking disease names is used for interlinking place names. URIs are

⁶² www.lsrn.org

normalized by appending “GeonamesID” field to the namespace of GeoNames⁶³ in order to be matchable with the URIs represented for GeoNames dataset. As HIV is a global priority Epidemic, our data representation is at a country level where users can search and visualize information for each country or they can aggregate by region or indicators. The developed tools for searching and visualization will be discussed in detail in the following chapter.

3.2.4. Statistical data Preparation

The HIV data about each African country was primarily extracted from the global health observatory dataset repository of world health organization. GHO data repository provides access to over 50 datasets on priority health topics including mortality and burden of diseases, the Millennium Development Goals (child nutrition, child health, maternal and reproductive health, immunization, HIV/AIDS, tuberculosis, malaria, neglected diseases, water and sanitation), non communicable diseases and risk factors, epidemic-prone diseases, health systems, environmental health, violence and injuries, equity among others.

Spreadsheet files, which were downloaded from GHO, contain both data and metadata. Typically, the rows and columns contain metadata, i.e., concepts (what the spreadsheet is about) and the center contains the data (instance values). During the publication process, it is very important to separate concepts from values, for concepts need to be represented using RDF vocabularies. The mapping of spreadsheet concepts to RDF vocabularies anchors the semantics of the triple set to be created. Transforming these spreadsheets to RDF in a fully automated way may cause information loss as there may be dimensions encoded in the heading or label of a sheet. [31]

⁶³ <http://sws.geonames.org/>

Most plugins for example Aperture ⁶⁴, POI ⁶⁵, RDFizers ⁶⁶ RDF123 ⁶⁷ or Stats2RDF ⁶⁸, provides support to the spreadsheet triplification. However, Most relational triplification tools maps tables to RDF classes and attributes to RDF properties, with no concern to identifying possible matches with existing standard vocabularies, thereby most often creating new vocabularies unnecessarily. [35] Thus we prefer a semi automatic method of triplification by using excel2rdf convertor and data cube vocabulary data structure.

3.2.5. Provenance Allocation

After producing the RDF triples, we include additional metadata and provenance information to facilitate retrieval and future interoperability of the published triple set. Provenance metadata is central to guarantee information reliability about the people, data sources, collection, and transformation processes the data went through. The plug-in currently supports the inclusion of the date of the triplification, file name used, the data source file, name of the user who performed the triplification, among others. In addition, Xcel2RDF also captures information on the vocabulary choices made during the triplification process. It is often the case that there is more than one choice of RDF vocabulary to be used in the process of annotating data. Registering the user choices is a very important step because it helps identify mappings to other vocabularies and schemas. The mappings are useful for data integration, and in the construction of mediators, that will consume the triple set in question. [35]

⁶⁴ <http://aperture.sourceforge.net/>)

⁶⁵ <http://poi.apache.org/>),

⁶⁶ <http://simile.mit.edu/wiki/RDFizers>)

⁶⁷ <http://rdf123.umbc.edu/>

⁶⁸ <http://aksw.org/Projects/Stats2RDF.html>

3.3. Data storage

Ontologies and other RDF data models can be stored in native RDF data stores or in relational databases that have been customized to support associative data techniques. Native RDF-data stores are inherently designed to support the concept of triples and can offer an efficient out-of-the-box approach to storing ontologies. RDF native databases are available from companies such as Tucana Technologies and Intellidimensions. Several high-quality open source RDF data stores also exist, including Kowari, Redland, Sesame, and 3Store. To use a relational database, the database must be designed in a somewhat non-traditional way. Instead of having a table that describes each major concept, the database design typically mimics the concept of triples by using a single table containing four columns. Three of the columns store the triple while the forth column is used to store its identification tag. [36]

Linked Data
storage technologies

Issues related to representing, storing, and querying using triples (i.e., RDF) versus traditional relational approaches, as well as the use and/or co-existence of the two types of data stores within implementations, are still working themselves out within industry and the marketplace. Each store-and-query facility provides unique capabilities that the other, at present, does not. RDF is great for situations when it is difficult to anticipate the types of queries that will be performed in the future. It is also terrific for handling metadata and for making queries that require inferences across imprecise or disparate data. [36]

"Triple Store" is the common name given to a database management system for RDF Data. These systems provide data management and data access via APIs and query languages to RDF Data. In actuality, many Triple Stores are in fact Quad Stores, due to the need to maintain RDF Data provenance within the data management system. Most Triple Store that supports Named Graph functionality are more likely a Quad Store. The most common known triple stores like Sesame, Virtuoso and Fuseki are all Quad stores. [13]

Fuseki is a SPARQL server. It provides REST-style SPARQL HTTP Update, SPARQL Query, and SPARQL Update using the SPARQL protocol over HTTP.

The screenshot shows the Fuseki web interface in a browser window. The address bar shows 'localhost:3030/sparql.tpl'. Below the browser window, the 'dataset: /lohd' is indicated. The 'SPARQL Query' section contains a text area with the following SPARQL query:

```
prefix geo: <http://www.w3.org/2003/01/geo/wgs84_pos#>
prefix dbpedia-owl: <http://dbpedia.org/ontology#>
prefix wgs84: <http://www.w3.org/2003/01/geo/wgs84_pos#>
prefix qb: <http://purl.org/linked-data/cube#>
prefix sdmx-measure: <http://purl.org/linked-data/sdmx/2009/measure#>

SELECT xsd:decimal(?lat) xsd:decimal(?lon) ?name ?url ?text ?url ?image
where { ?lohd wgs84:lat ?lat; wgs84:long ?lon; geo:name ?name. optional { ?lohd
rdfs:isDefinedBy ?url; geo:image ?image; loh:label ?text; geo:image . } }
```

Below the query text area, there is an 'Output:' dropdown menu set to 'XML', an 'XSLT style sheet (blank for none):' text box containing '/xml-to-html.xsl', a checkbox for 'Force the accept header to text/plain regardless' which is unchecked, and a 'Get Results' button.

The 'SPARQL Update' section contains a text area with the query 'DROP DEFAULT' and a 'Perform update' button.

The 'File upload' section contains a 'File:' label, a 'Choose Files' button, a file icon, and the filename '1990final.ttl'. Below this is a 'Graph:' label and a dropdown menu set to 'default'.

Figure 3 : FUSEKI Triple store Interface

It is important to note that RDF query languages are still evolving, which may to some extent explain this limitation. Other limitations of RDF relate to performance issues. Because queries can be broadened, for example, to include concepts instead of just terms, the search space can be dramatically increased. Because RDF data stores are relatively new and the number of implementations relatively small, system developers need to iterate over their designs, paying particular attention to queries and functions that could

have negative effects on performance. In terms of industry growth, it is difficult to predict how RDF will affect the database industry. RDF data stores may remain a distinct data storage category in their own right or their capabilities may be subsumed into relational databases in a manner similar to what happened with object-oriented databases. [13]

3.4. Data License

All our published Linked Data adheres to original data publisher's data license and terms of use. Additionally attributions are given for each source accordingly.

3.5. Data Enrichment

In this section of the thesis we will discuss about some of the ways that we tried to enrich the original data by adding information such that the datasets can be more useful, better discovered, interlinked or queried. The primary motivation in representing health data using linked open data principle is to be able to link health data which are from different sources and to discover and use them in our data analysis which will be discussed in detail in the next section.

3.6. Data Interlinking

The Semantic Web isn't just about putting data on the web. It is about making links, so that a person or machine can explore the web of data. With linked data, when you have some of it, you can find other related data. [5] The Web of Data is built upon two simple ideas: First, to employ the RDF data model to publish structured data on the Web. Second, to set explicit RDF links⁶⁹ between data items within different data sources. [37]

⁶⁹ <http://wifo5-03.informatik.uni-mannheim.de/bizer/pub/LinkedDataTutorial/#links>

Interlinking things and concepts from our RDF datasets to external datasets in the LOD Cloud was a challenging task. It primarily requires an investigation to identify eligible resource types in our datasets, and then finding suitable matching resources in external datasets. One requirement was to make sure that the target resources were dereferenceable in order to make the interlinking worthwhile. Tables [1], [2], give an overview of the targeted external datasets, entity types, links, and counts for each of the three cases respectively.

In our data enrichment, we use both manual and automatic ways of data enrichment. For some of the data, which are difficult and inconsistent, to be discovered we use the manual method and for some of the data, which have a lot of links, which make handling it difficult, we use the automatic method of data discovery and linkage.

Silk framework

The Silk Link Discovery Framework⁷⁰ consists of a console application used to interlink two data sets as well as of the Silk Server, an HTTP server, which receives an incoming RDF stream and creates links between data items. Both applications provide a flexible configuration language, the Silk Link Specification Language (Silk-LSL), to specify the conditions data items must fulfill in order to be interlinked. [38]

For accessing the source and target data sources, we first configure access parameters to the target dataset endpoints using the **<DataSource>** directive. The only mandatory data source parameter is the endpoint URI. By specifying the source and destination endpoints on target datasets, we make an interlinking of the following data.

⁷⁰ <http://wifo5-03.informatik.uni-mannheim.de/bizer/silk/>

Target Dataset	Link type	Link count
DBpedia	Owl:SameAs	47
Geonames	Owl:SameAs	46
diseasome	Owl:CloseMatch	11
World Bank	qb:GDP	94

Table 1: Manual Interlinking Results

Target Dataset	Instances	Link type	Link count
PubMed	618	Owl:SameAs	311
LinkedCT	213	Owl:SameAs	111
BIO2rdf	114	Owl:SameAs	97

Table 2: Automatic Interlinking results

Overall steps taken in methodic background section and methodology of the data management chapter are depicted in the following workflow diagram at Figure 4. Explanation of Live visualization and interfaces for query writing as well as the over all system architecture using layers is explained in detail in the following Chapter. Additionally, we had documented some of the visualization and search results of the system.

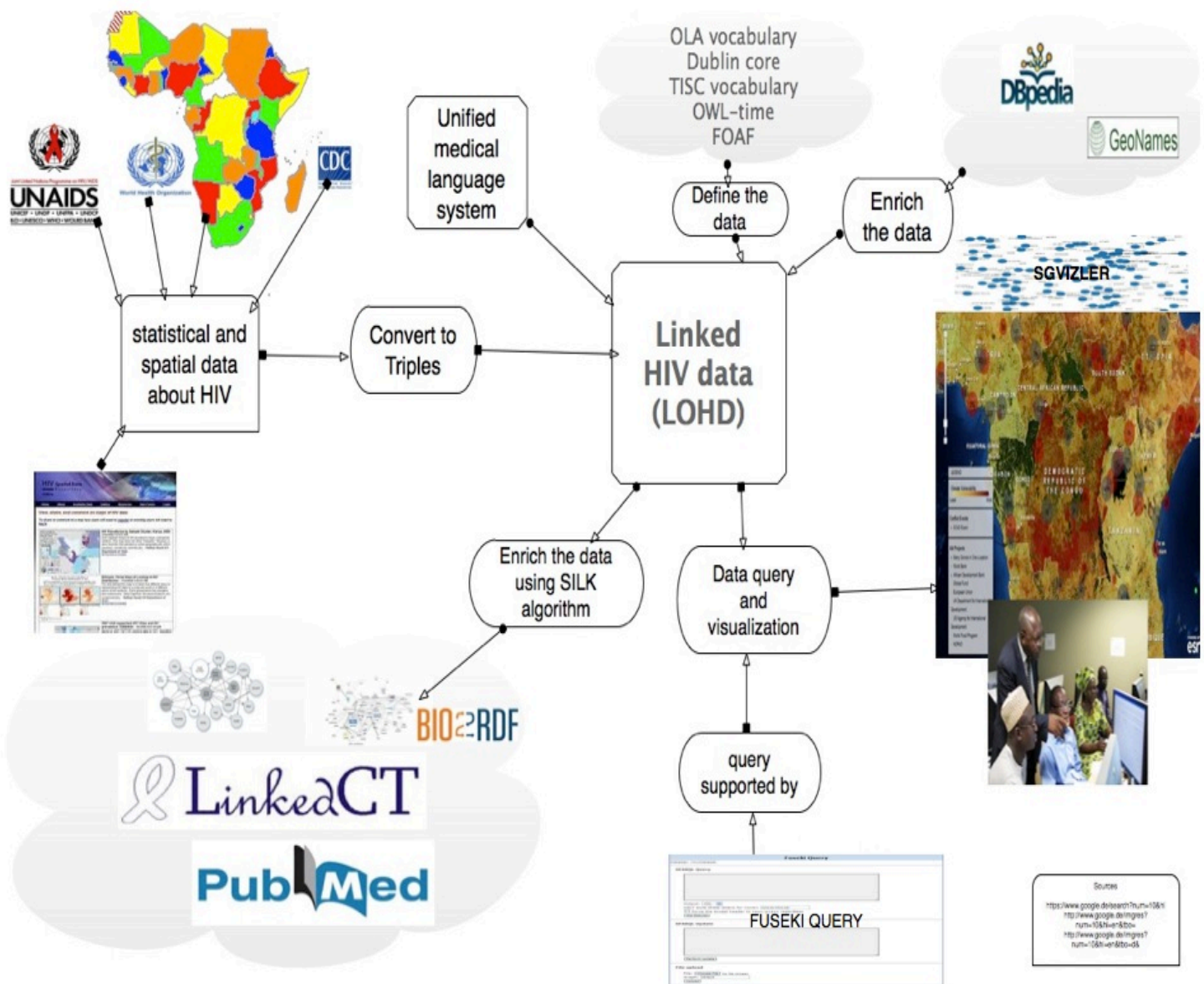


Figure 4: Workflow Diagram of Methodology

4. System Overview

4.1. General overview

The Linked Open Health Data (LOHD) platform, which is a domain specific application, uses RDF data stored on Fuseki triple store. It provides a semantic platform for Health data that integrates data from a set of distributed open official data sources including geographical and statistical health data from various sources. In the system, users can query HIV related information about African countries and the system will support them in querying and visualizing the data both in space and time.

Linked Data based Health Information System on the Semantic Web

- [Home](#)
- [Data](#)
- [Queries](#)
- [Visualizations](#)

- HIV trend visualization
 - [East Africa](#)
 - [Westm africa](#)
 - [Central Africa](#)
 - [Northern Africa](#)
 - [Southern Africa](#)

User Input

This section contains a input form where users can write and execute their own SPARQL queries. The query is sent to the Fuseki triple store where all the converted Who data is stored via the URL in GET parameters.

```
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 xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX lohd: <http://localhost:3030/lohd#>
PREFIX loh: <http://localhost:3030/loh#>
PREFIX dbpedia: <http://www.dbpedia.org/ontology/>
PREFIX geo: <http://www.w3.org/2003/01/geo/wg84#>
PREFIX dbpedia-owl: <http://dbpedia.org/ontology/>
PREFIX wgs84: <http://www.w3.org/2003/01/geo/wg84#>
PREFIX qb: <http://purl.org/linked-data/query#>

SELECT xsd:decimal(?lat) xsd:decimal(?lon)
WHERE {
  ?image lohd:isDef ?url ?text ?url
  ?name optional { ?lohd rdfs:isDef ?image ?image; loh:label ?name }
}
```

gBarChart

gSparkline

gScatterChart

gCandlestickChart

gOrgChart

gTimeline

gMotionChart

gGeoChart

gGeoMap

gVMap

gTable

dForceGraph

rdGraph

sDefList

sList

sMap

sTable

sText

Width: 800 Height: 400 Chart Type: sMap Reset GO!

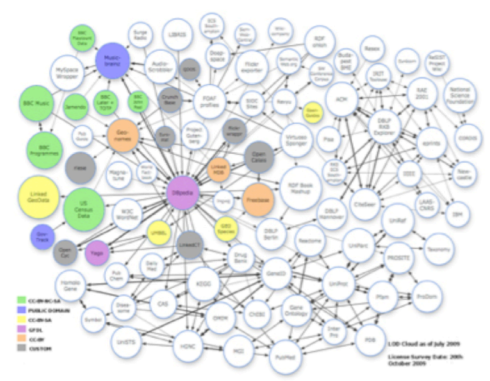


Figure 5: The home page of LOHD system

4.2. LOHD System Architecture

LOHD is a system built up on different layers. Multi layer architecture provides flexibility and reusability in which data management, Query processing and Visualization 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 hierarchy, different layers can be developed sequentially and modified asynchronously without affecting the entire system Architecture. [13] Multi layer architecture is composed of 4 main layers. Data layer, which consists of various data servers and formats that communicate to presentation layer through an intermediary layer composed by application query and visualization services.

Depicting information systems in a layered approach has its own advantages. Hartig and Langegger present a deeper comparison of the advantages and disadvantages of the architectural patterns [39]. The appropriate pattern (or mixture of these patterns) will always depend on the specific needs of a Linked Data application⁷¹ we want to build. Considering the heterogeneity of the data sources, the required response time and the extent of data discovery [40] we need, to model our data with a layered approach.

Layered approach
advantages

However, due to the likelihood of scalability problems with on-the-fly link traversal and federated querying, applying new and additional service layers may transpire that widespread crawling and caching will become the norm in making data from a large number of data sources available to applications with acceptable query response times, while being able to take advantage of the openness of the Web of Data by discovering new data sources through link traversal.

⁷¹ <http://www.semanticweb.gr/topos/>

The overall system architecture's layers are explained in the following figure including all the components in the system.

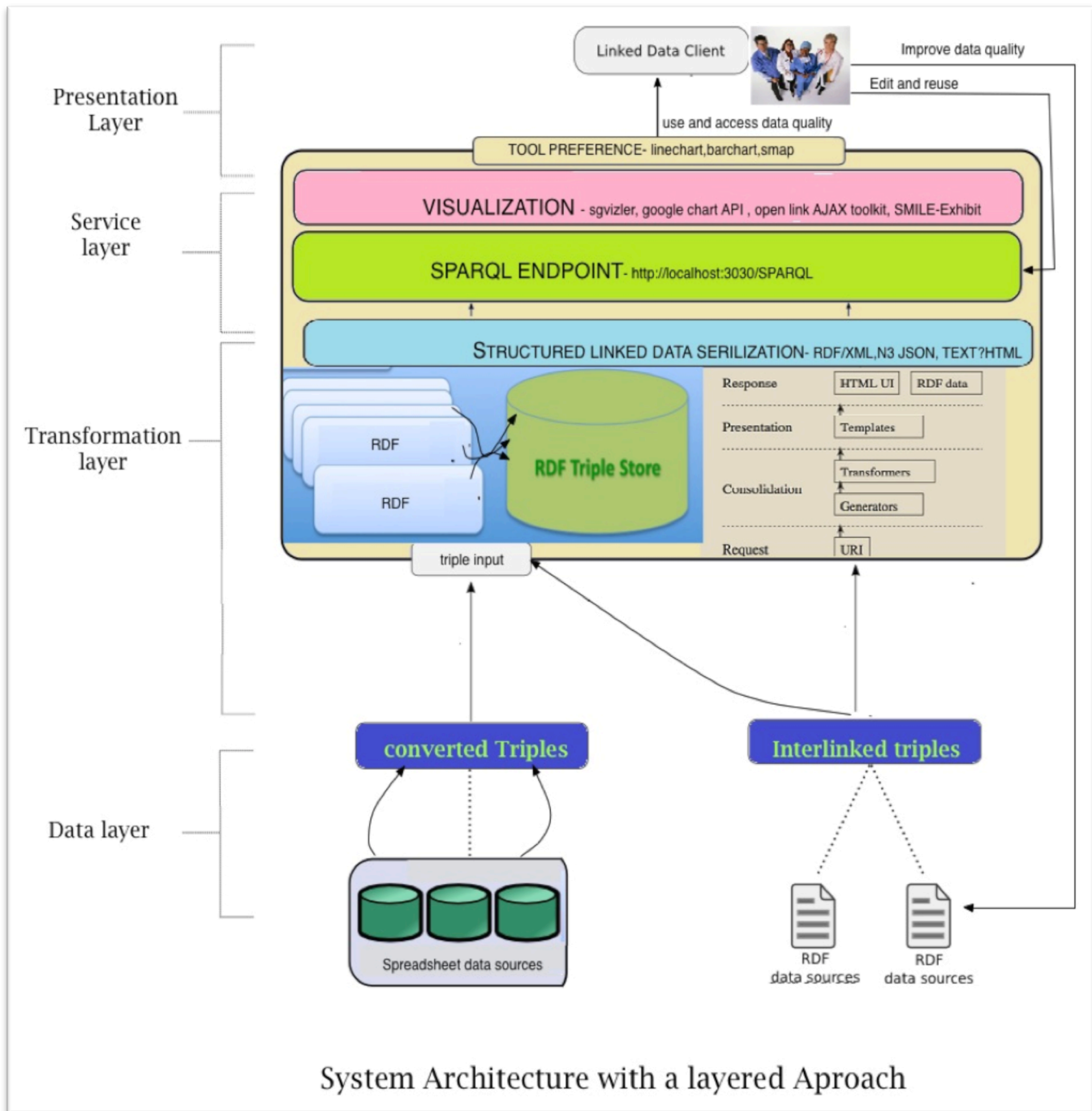


Figure 6: LOHD System architecture with a layered approach

4.2.1. Data Layer

Data layer is the first layer where all the converted and interlinked data is stored and managed. All HIV related data was retrieved from WHO for 20 years from 1990 to 2010 and to be able to do time series and trend analysis, missing information was complemented by other sources. All this data was retrieved in excel format and was converted to RDF. Other related statistical and geographical data that are already on the cloud was interlinked using both manual and automatic method methods using SILK link discovery framework. All this data was imported in to Fuseki triple store. Fuseki triple store offers more than a RDF store with promising capabilities for solving data silos problem by providing data in different formats using a local end point from various departments and disparate data sources in several formats, by delivering an unrivaled platform for real-time access, integration querying and visualization. Currently the system have more than 21,000 triples in its dataset and when this content becomes available on the web, anyone interested can reuse, visualize with the built in sgvizler visualization tools and query the system. The format of piece of data is shown in the following figure.

Data layer triple
count

Subject	Predicate	Object
<http://localhost:3030/lohd/data#region_101>	<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>	<http://www.w3.org/2000/01/rdf-schema#Class>
<http://localhost:3030/lohd/data#region_101>	<http://localhost:3030/lohd/data#dataSource>	* <http://apps.who.int/gho/data/> *
<http://localhost:3030/lohd/data#region_101>	<http://localhost:3030/lohd/data#hasObservation>	"HIV Prevalence in 1990"
<http://localhost:3030/lohd/data#region_101>	<http://localhost:3030/lohd/data#hasObservation>	"HIV Prevalence in 1991"
<http://localhost:3030/lohd/data#region_101>	<http://localhost:3030/lohd/data#hasObservation>	"HIV Prevalence in 1992"
<http://localhost:3030/lohd/data#region_101>	<http://localhost:3030/lohd/data#hasObservation>	"HIV Prevalence in 2000"
<http://localhost:3030/lohd/data#region_101>	<http://localhost:3030/lohd/data#hasObservation>	"HIV Prevalence in 1996"
<http://localhost:3030/lohd/data#region_101>	<http://localhost:3030/lohd/data#hasObservation>	"HIV Prevalence in 1999"
<http://localhost:3030/lohd/data#region_101>	<http://localhost:3030/lohd/data#hasObservation>	"HIV Prevalence in 1997"
<http://localhost:3030/lohd/data#region_101>	<http://localhost:3030/lohd/data#hasObservation>	"HIV Prevalence in 1993"
<http://localhost:3030/lohd/data#region_101>	<http://localhost:3030/lohd/data#hasObservation>	"HIV Prevalence in 1998"
<http://localhost:3030/lohd/data#region_101>	<http://localhost:3030/lohd/data#hasObservation>	"HIV Prevalence in 1995"
<http://localhost:3030/lohd/data#region_101>	<http://localhost:3030/lohd/data#hasObservation>	"HIV Prevalence in 1994"
<http://localhost:3030/lohd/data#region_101>	<http://www.w3.org/2003/01/geo/wgs84_pos#nearbyFeatures>	*<http://www.geonames.org/337996/nearby>*
<http://localhost:3030/lohd/data#region_101>	<http://www.w3.org/2003/01/geo/wgs84_pos#alternateName>	"ኢትዮጵያ"
<http://localhost:3030/lohd/data#region_101>	<http://www.w3.org/2003/01/geo/wgs84_pos#long>	"40.12207"
<http://localhost:3030/lohd/data#region_101>	<http://www.w3.org/2000/01/rdf-schema#isDefinedBy>	*<http://www.geonames.org/337996/about.rdf>*
<http://localhost:3030/lohd/data#region_101>	<http://www.w3.org/2000/01/rdf-schema#label>	"Ethiopia "
<http://localhost:3030/lohd/data#region_101>	<http://localhost:3030/lohd/data#HIV_Prevalence_in_1990>	"0.7"
<http://localhost:3030/lohd/data#region_101>	<http://www.w3.org/2000/01/rdf-schema#comment>	"Prevalence of HIV in Ethiopia in the year 1990"
<http://localhost:3030/lohd/data#region_101>	<http://www.w3.org/2000/01/rdf-schema#comment>	"Prevalence of HIV in Ethiopia in the year 1991"
<http://localhost:3030/lohd/data#region_101>	<http://www.w3.org/2000/01/rdf-schema#comment>	"Prevalence of HIV in Ethiopia in the year 1993"
<http://localhost:3030/lohd/data#region_101>	<http://www.w3.org/2000/01/rdf-schema#comment>	"Prevalence of HIV in Ethiopia in the year 1994"
<http://localhost:3030/lohd/data#region_101>	<http://www.w3.org/2000/01/rdf-schema#comment>	"Prevalence of HIV in Ethiopia in the year 1992"
<http://localhost:3030/lohd/data#region_101>	<http://www.w3.org/2000/01/rdf-schema#comment>	"Prevalence of HIV in Ethiopia in the year 1997"
<http://localhost:3030/lohd/data#region_101>	<http://www.w3.org/2000/01/rdf-schema#comment>	"Prevalence of HIV in Ethiopia in the year 1999"

Table 2: Sample triple data in the Fuseki triple store data layer

4.2.2. Transformation layer

The architectures of Linked Data applications are very diverse and largely depend on the concrete use case. In our case, the transformation layer is a layer that helps to bridge the data and the service layer. It is the layer where every SPARQL query get processed using crawling pattern.⁷² Applications that implement this pattern crawl the Web of Data in advance by traversing RDF links. Afterwards, they integrate and cleanse the discovered data and provide the higher layers of the application with an integrated view on the original data.

Transformation layer employs the R2R Framework⁷³ to translate Web data that is represented using terms from different vocabularies into a single target vocabulary. Vocabulary mappings are expressed using the R2R Mapping Language.⁷⁴ The language provides for simple transformations as well as for more complex structural transformations and property value transformations such as normalizing different units of measurement or complex string manipulations. The syntax of the R2R Mapping Language is very similar to the query language SPARQL, which eases the learning curve. The expressivity of the language enabled us to deal with all requirements that we have encountered so far when translating Linked Data from the Web into a target representation. [40]

Transformation
layer mapping

The transformation layer finds different URIs that are used within different data sources to identify the same real-world entity. For each set of duplicates that have been identified by Silk, fuseki replaces all URI aliases with a single target URI within the output data. In addition, it adds owl:sameAs links pointing at the original URIs, which makes it possible for applications to refer back to the data sources on the Web. If the triple input data already contains owl:sameAs links, the referenced URIs are normalized accordingly.

⁷² <http://linkeddatabook.com/editions/1.0/>

⁷³ <http://wifo5-03.informatik.uni-mannheim.de/bizer/r2r/>

⁷⁴ <http://wifo5-03.informatik.uni-mannheim.de/bizer/r2r/spec/>

The crawling pattern mimics the architecture of classical Web search engines like Google and Yahoo. The crawling pattern is suitable for implementing applications on top of an open, growing set of sources, as the crawler at run-time discovers new sources. Separating the tasks of building up the cache and using this cache later in the application context enables applications to execute complex queries with reasonable performance over large amounts of data as depicted in the following figure. [38]

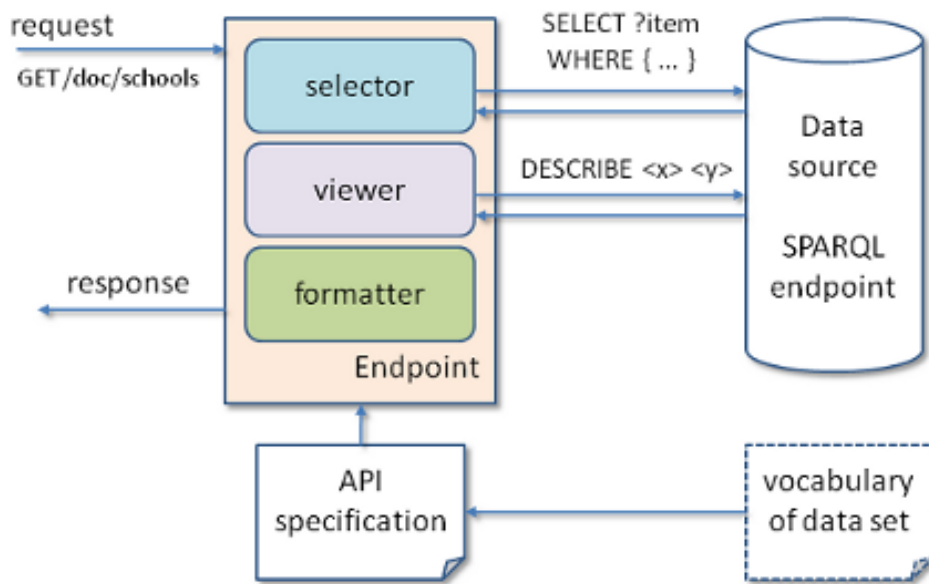


Fig 7: Query processing in the transformation layer⁷⁵

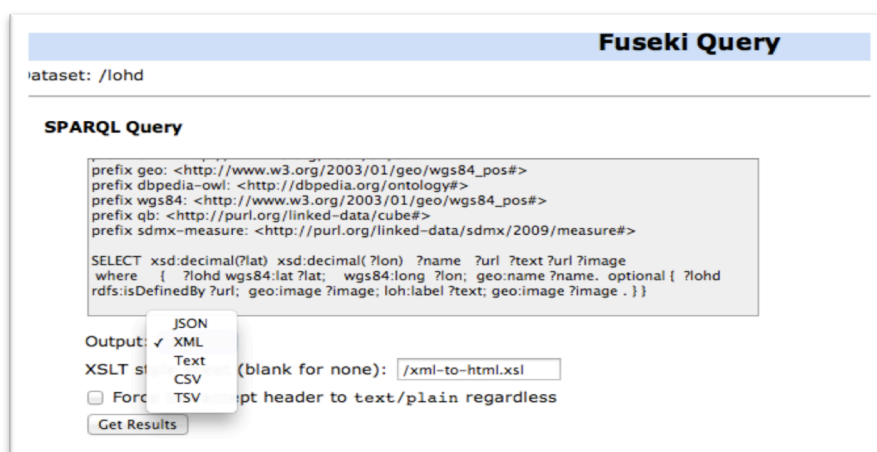
4.2.3. Service layer

This is the backbone intermediary layer of the system, which basically controls the data access and bridges the client to the server via services by processing the commands of the clients. Most of the service request is through SPARQL queries that are explained in the following section.

⁷⁵ Adapted from <http://linkeddatabook.com/editions/1.0/>

4.2.3.1. SPARQL endpoint

A **SPARQL** endpoint is a conformant SPARQL protocol service as defined in the **SPROT**⁷⁶ specification. A SPARQL endpoint enables users (human or other) to query a knowledge base via the SPARQL language. Results are typically returned in one or more machine-processable formats. Therefore, a SPARQL endpoint is mostly conceived as a machine-friendly interface towards a knowledge base. Both the formulation of the queries and the human-readable presentation of the results should typically be implemented by the calling software, and not be done manually by human users. [38] In our system, the SPARQL interface of Fuseki, allows to write SPARQL query, update data and/or to upload data files as shown in the following figure.



The screenshot shows the 'Fuseki Query' web interface. At the top, there is a header 'Fuseki Query' in a blue bar. Below it, a text field contains 'dataset: /lohd'. The main section is titled 'SPARQL Query' and contains a text area with the following SPARQL query:

```
prefix geo: <http://www.w3.org/2003/01/geo/wgs84_pos#>
prefix dbpedia-owl: <http://dbpedia.org/ontology#>
prefix wgs84: <http://www.w3.org/2003/01/geo/wgs84_pos#>
prefix qb: <http://purl.org/linked-data/cube#>
prefix sdmx-measure: <http://purl.org/linked-data/sdmx/2009/measure#>

SELECT xsd:decimal(?lat) xsd:decimal(?lon) ?name ?url ?text ?url ?image
WHERE { ?lohd wgs84:lat ?lat; wgs84:long ?lon; geo:name ?name. optional { ?lohd
rdfs:isDefinedBy ?url; geo:image ?image; lohd:label ?text; geo:image ?image . } }
```

Below the query text area, there are output options. A dropdown menu is open, showing 'JSON', 'XML' (selected), 'Text', 'CSV', and 'TSV'. To the right of the dropdown, there is a text field '(blank for none): /xml-to-html.xsl'. Below the dropdown, there is a checkbox 'Force TSV' and a checkbox 'Accept header to text/plain regardless'. At the bottom left, there is a 'Get Results' button.

Fig 8: Sample SPARQL Query

In the LOHD system, the enriched HIV data about each Sub Saharan African countries can be accessed via the SPARQL interface of the triple store. SPARQL endpoint serves the data like a proxy service to Linked Data frontends. SPARQL endpoint can be accessed directly through 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,

⁷⁶ <http://semanticweb.org/wiki/SPROT>

timeline and visualizations. Triples stored in the Fuseki triple store, information encoded using reification and the provenance of the assertions can be queried with SPARQL queries. The realization of advanced queries and testing for advanced visualization is still a further research work.

4.2.4. Presentation Layer

Presentation layer is the upper level layer 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 different RDF stores support different querying and visualization tools. In our system, we embedded methods both for visualization and information searching. Users who are able to write SPARQL queries can write different queries and they can choose a kind of visualization tool they need. Other end users specifically health professionals and healthcare managers who are not aware of SPARQL queries can use the linked data search engine to search for the information he want just by giving a simple text like a search on Google.

In the following section we will discuss the tools, which are available for visualization and interfaces for searching like other search engines. Visualizations through SPARQL queries can give advanced results for users. The development of search engine interface is still on beta phase and currently the user is advised to use the SPARQL interface for better result.

4.2.4.1. Visualization

In the last years, the amount of semantic data available on the Web has increased dramatically, especially thanks to initiatives like Linked Open Data (LOD). The potential of this vast amount of data is enormous but in most cases it is very difficult and cumbersome for users to visualize, explore and use this data, especially for lay-users without experience with Semantic Web technologies. [44]

The goal of information visualization is to translate abstract information into a visual form that provides new insight about that information. The uptake and consumption of Linked Data is currently restricted almost entirely to the Semantic Web community. While the utility of Linked Data to non-tech savvy web users is evident, the lack of technical knowledge and an understanding of the intricacies of the semantic technology stack limit such users in their ability to interpret and make use of the Web of Data. [43] A key solution in overcoming this hurdle is to visualize Linked Data in a coherent and legible manner, allowing non-domain and non-technical audiences to obtain a good understanding of its structure, and therefore implicitly compose queries, identify links between resources and intuitively discover new pieces of information which are useful for decision making.

The size and scale of the Web of Data presents challenges when trying to make sense of the information contained within it. A basic visualization of the Web of Data, focusing on a resource, which has a high outdegree of relationships to other data, will present the viewer with a mass of edges linking into the resource, resulting in information overload. How does an end user make sense of the response? How do they understand and interpret the data in a meaningful way? The third principle of Linked Data states that “When someone looks up a URI, provide useful information, using the standards ” [5]; therefore when a URI (UniformResource Identifier) is dereferenced, a response is returned according to the requester’s parameters. These parameters can request an XHTML⁷⁷ (the eXtensible HyperText Markup Language) [41] representation of the resource – in which case the information can be displayed in a Web browser, while embedding machine-readable information in RDF according to a given serialization format. In the latter case knowledge of how to use this format and interpret the information provided using it is restricted to tech-savvy end users, and in certain cases, only those who have knowledge of Semantic Web (SW) technologies. It is clear that regular (readable) Web users, so called lay users,

Challenges in
Visualization

⁷⁷ <http://www.w3.org/TR/xhtml1/>

who have no knowledge of RDF, nor ontologies, are inhibited in their ability to understand data returned when looking up a URI [24,29,42].

Clear and coherent visualization of Linked Data would enable accessibility to the Web of Data and encourage its use outside the SW community. To enable such uptake therefore requires Linked Data to become usable also by lay users, by providing interfaces and browsers of the Web of Data to support sense making and information exploration and discovery. Furthermore, query composition in languages such as SPARQL [6], although useful, requires understanding of a given query language's syntax and at least a basic knowledge of data content and structure. End users should be able to implicitly compose such queries without being aware of the underlying query mechanism that is used to pose the required questions. To overcome those challenges we have proposed two solutions. The first one is to develop a visualization interface and the second one is to develop an interface for search engines that can crawl over the data with simple search strings from lay users. In the next section we will show how we handle the visualization using SGVIZLER visualization tool and the search engine tool.

4.3.4.1.1. Sgvizler

Sgvizler⁷⁸(Sgvizler) is a javascript which renders the result of SPARQL SELECT queries into charts or html elements. The name and tool relies on and/or is inspired by SPARQL, Google Visualization API⁷⁹ (Google, Google Visualization API Reference), SPARQLer Snorql (Snorql)⁸⁰and Spark⁸¹. All the major chart types offered by the Google Visualization API are supported by Sgvizler. The user inputs a SPARQL query which is sent to a designated SPARQL endpoint. The endpoint must return the result with the selected visualization type. The user inputs a SPARQL query which is sent to a

⁷⁸ <http://code.google.com/p/sgvizler/>

⁷⁹ <https://developers.google.com/chart/interactive/docs/reference>

⁸⁰ <http://data.semanticweb.org/snorql/>

⁸¹ <http://www.novospark.com/>

designated SPARQL endpoint. The endpoint returns the results back in SPARQL Query Results XML Format or SPARQL Query Results in JSON format. Sgvizler parses the results into the JSON format that Google prefers and displays the chart using the Google Visualization API or a custom-made visualization or formatting function. [43].

When designing SPARQL queries for visualization by Sgvizler, the order of the columns in the result set, i.e., the order of the variables in the SELECT block, is crucial. It expects values and requests to be in order. Regarding SPARQL endpoint communication, Sgvizler has the same web browser compatibility as the external JavaScript libraries it uses. [43] For JQuery and the Google Chart Tools this means compatibility with all reasonably new web browsers.⁸⁵

SGVIZLER provides more than twenty types of visualization tools that visualize the SPARQL select results. The main point in the tools is that, the tools are dependent on the data type of the SPARQL return query. If it is a continues data you can visualize them using either `glinechart`⁸², `gareachart`⁸³, `gsparkline`⁸⁴ or `gtimeline`⁸⁵ if it is a data with consecutive years. In the following figure, we display the different visualizations we make on the system. For more queries and live visualizations please visit www.observchange.com/lohd where we have published all the codes and results.

⁸² <http://sgvizler.googlecode.com/svn/www/screenshots/gLineChart.png>

⁸³ <http://sgvizler.googlecode.com/svn/www/screenshots/gAreaChart.png>

⁸⁴ <http://sgvizler.googlecode.com/svn/www/screenshots/gAreaChart.png>

⁸⁵ <http://sgvizler.googlecode.com/svn/www/screenshots/gTimeline.png>

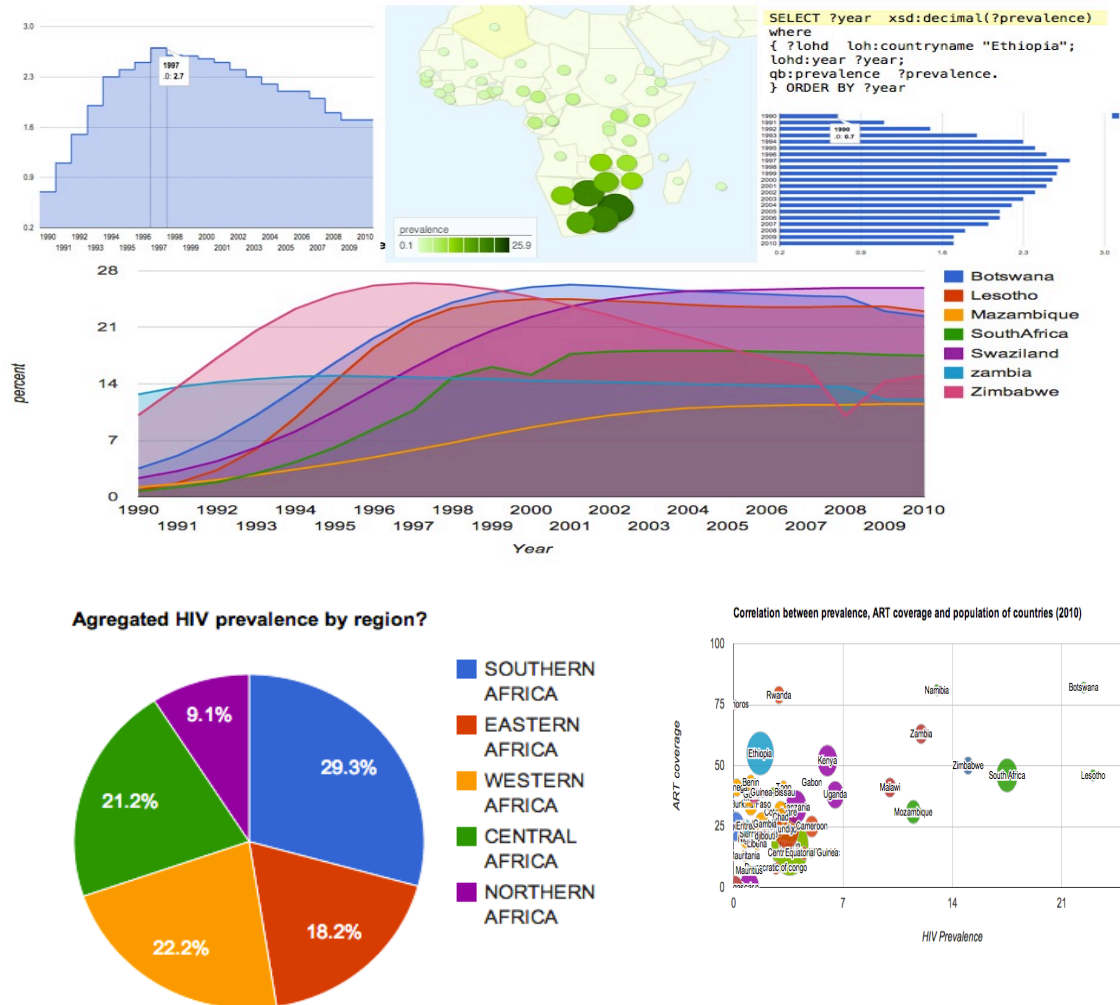


Figure 9: Sample visualizations using SGVIZLER over the LOHD data

All the above are visualizations of SPARQL results from the data which is stored on our system. All the SPARQL results and additional visualizations will be published online. For further information how to use the visualization tools, SGVIZLER have a nice online tutorial, <http://code.google.com/p/sgvizler/wiki/Introduction>⁸⁶.

⁸⁶ <http://code.google.com/p/sgvizler/wiki/Introduction>

4.3. Linked data search Engines

Just as traditional Web browsers allow users to navigate between HTML pages by following hypertext links, Linked Data browsers allow users to navigate between data sources by following RDF links. For example, a user may view LOHD's RDF description of the Burdon of HIV in Ethiopia, follow a population link to the description of Ethiopian population pattern (which originated from the interlinked dbpedia data), and from there onward into RDF data about the ART coverage in that country and the related efforts information. The result is that a user may begin navigation in one data source and progressively traverse the Web by following RDF rather than HTML links.

Linked data have been a priority research agenda in the semantic web community in the last years. [5] However, there is still lack of applications with good user interfaces to make Linked Data resources accessible to end-users. To address this address this there are several ongoing institutional and W3C endeavors.

A number of search engines have been developed that crawl Linked Data from the Web by following RDF links, and provide query capabilities over aggregated data. These search engines integrate data from thousands of data sources and thus nicely demonstrate the advantages of the open, standards-based Linked Data architecture, compared to Web 2.0 mashups which rely on a fixed set of data sources exposing proprietary interfaces. [44]

Search engines such as Sig.ma [44], provide keyword-based search services oriented towards human users and follow a similar interaction paradigm as existing market leaders such as Google and Yahoo. The user is presented with a search box into which they can enter keywords related to the item or topic in which they are interested, and the application returns a list of results that may be relevant to the query.

However, rather than simply providing links from search results through to the source documents in which the queried keywords are mentioned, Linked Data search engines provide richer interaction capabilities to the user that exploit the underlying structure of the data.

The Sig.ma search engine applies vocabulary mappings to integrate Web data as well as specific display templates to properly render data for human consumption. Figure 10 shows the Sig.ma search engine displaying data about HIV in South Africa that has been integrated from 17 data sources. Another interesting aspect of the Sig.ma search engine is that it approaches the data quality challenges that arise in the open environment of the Web by enabling its users to choose the data sources from which the user's aggregated view is constructed. By removing low quality data from their individual views, Sig.ma users collectively create ratings for data sources on the Web as a whole.

In our system we had integrated this interface and users who are not familiar with SPARQL queries can use this search engine interface to search the information they want.

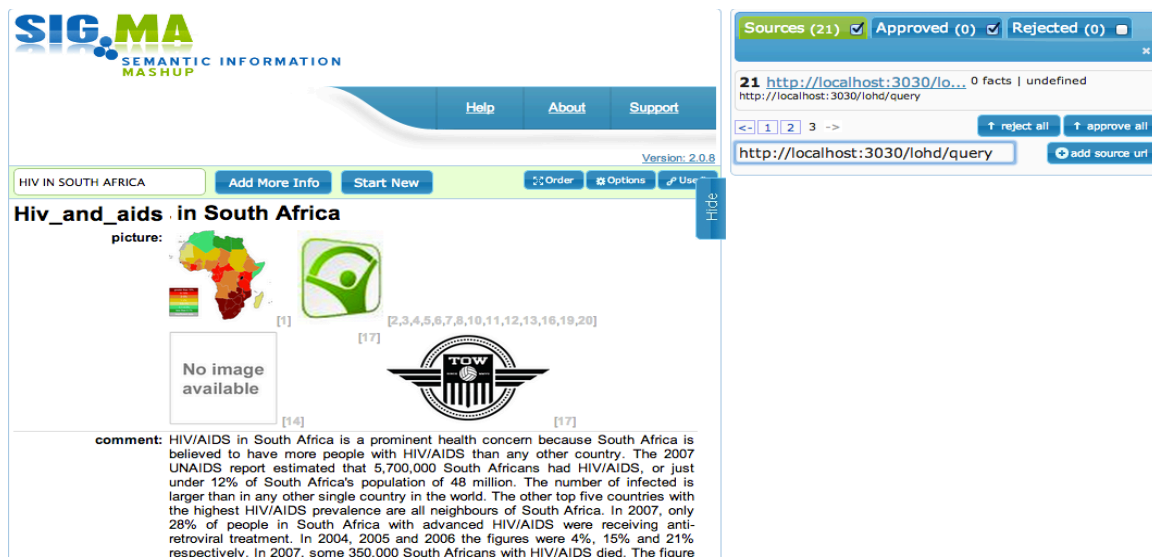


Figure 10: sig.ma search engine over LOHD data

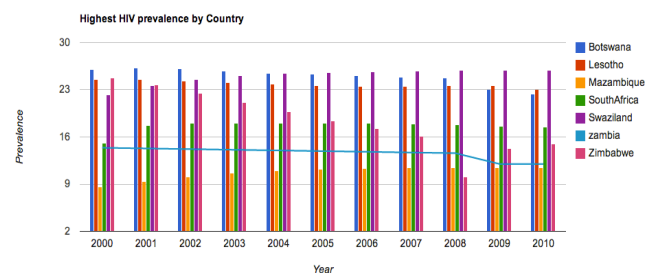
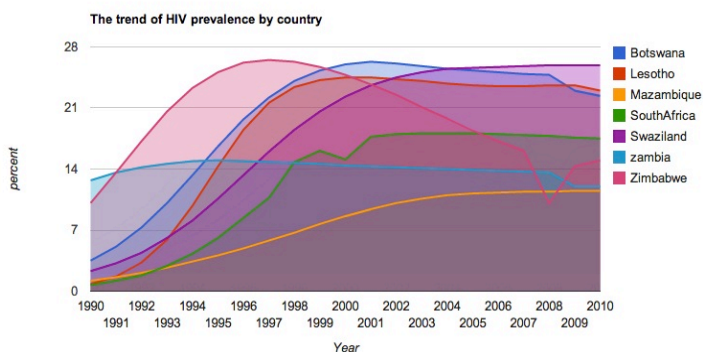
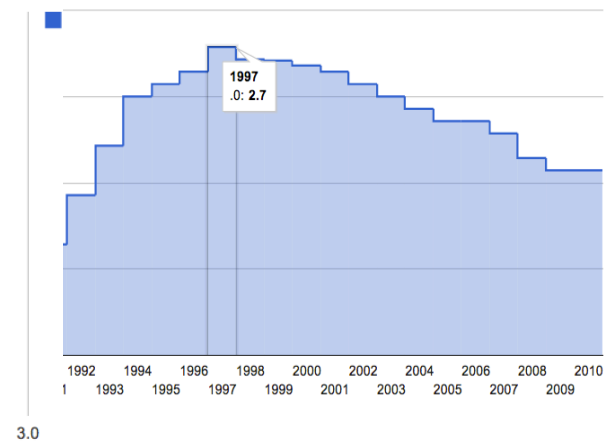
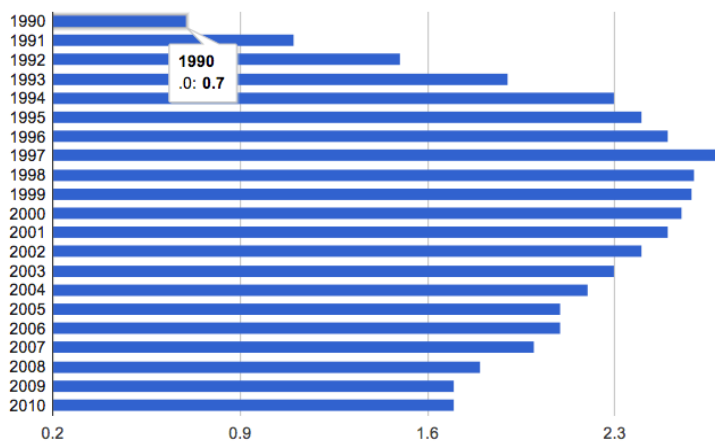
Currently search engine tools of Linked Data are still limited in the capability of querying. However, we explain how the machine can use the data and the vocabularies in browsing through the dataset to obtain the query results. The tool, sig.ma, has a big potential in searching resources but still it have some limitations in supporting advanced searching for users.

4.4. Case Studies of queries and visualizations

1. Time serious visualization over linked data

```
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 xsd: <http://www.w3.org/2001/XMLSchema#> PREFIX
lohd: <http://localhost:3030/lohd/data#> PREFIX loh:
<http://localhost:3030/lohd/d#> PREFIX dbpedia:
<http://www.dbpedia.org/resource#> PREFIX geo:
<http://www.w3.org/2003/01/geo/wgs84_pos#> PREFIX
dbpedia-owl: <http://dbpedia.org/ontology#> PREFIX
wgs84: <http://www.w3.org/2003/01/geo/wgs84_pos#> PREFIX
qb: http://purl.org/linked-data/cube#
```

```
SELECT ?year xsd:decimal(?prevalence)
where
{
  ?lohd lohd:year ?year;
    loh:countryname "Ethiopia";
    qb:prevalence ?prevalence.
}
order by ?year
```



2. Geographical visualization over linked data

```
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 xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX loh: <http://localhost:3030/loh/data#>
PREFIX loh: <http://localhost:3030/loh/d#>
PREFIX dbpedia: <http://www.dbpedia.org/resource#>
PREFIX geo: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX dbpedia-owl: <http://dbpedia.org/ontology#>
PREFIX wgs84: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX qb: http://purl.org/linked-data/cube#
```

```
SELECT xsd:decimal(?lat) xsd:decimal(?lon) ?name
?url ?text ?url ?image
```

where

```
{
  ?loh wgs84:lat ?lat;
  wgs84:long ?lon;
  geo:name ?name.

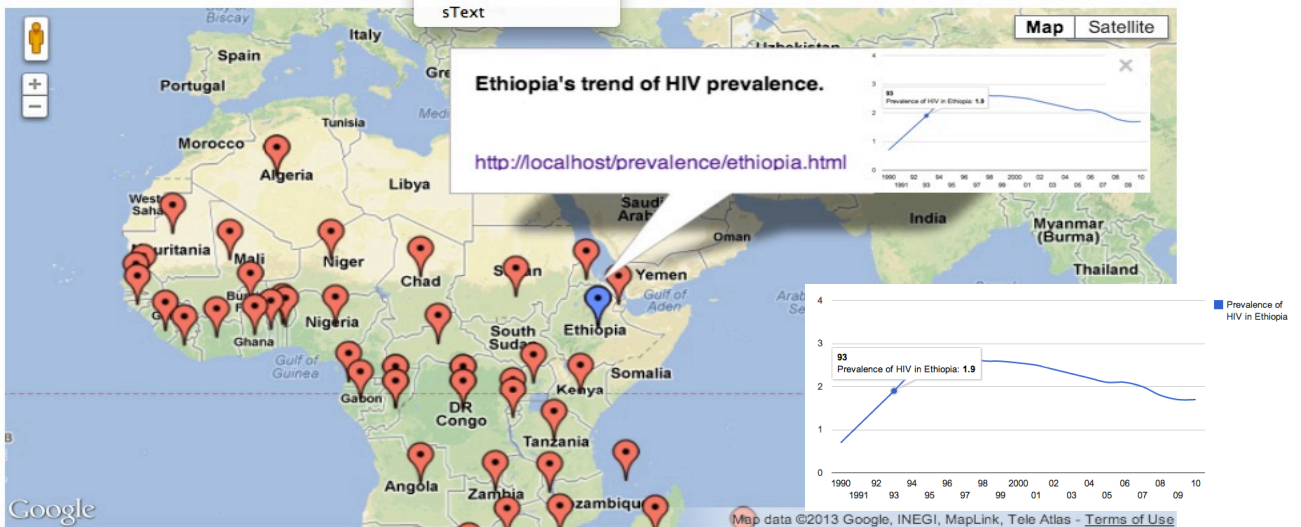
  optional { ?loh rdfs:isDefinedBy ?url; geo:image ?image;

  loh:label ?text; geo:image ?image . }
}
```

Width: Height: Chart Type

Received 58 rows. Drawing chart...
[View query results](#) (in new window).

- ☐ gMap
- ☐ gTable
- ☐ dForceGraph
- ☐ rdGraph
- ☐ sDefList
- ☐ sList
- ☒ sMap
- ☐ sTable
- ☐ sText



3. Indicator based visualization over linked

Additionally, we can make queries on the system on different indicators, like HIV prevalence rate by country or region, ART coverage rate, population and GDP and make a correlation analysis between those variables over time. An advanced correlation analysis is left for future work but in the following figures we show how those indicators can affect each other over a trend of time.

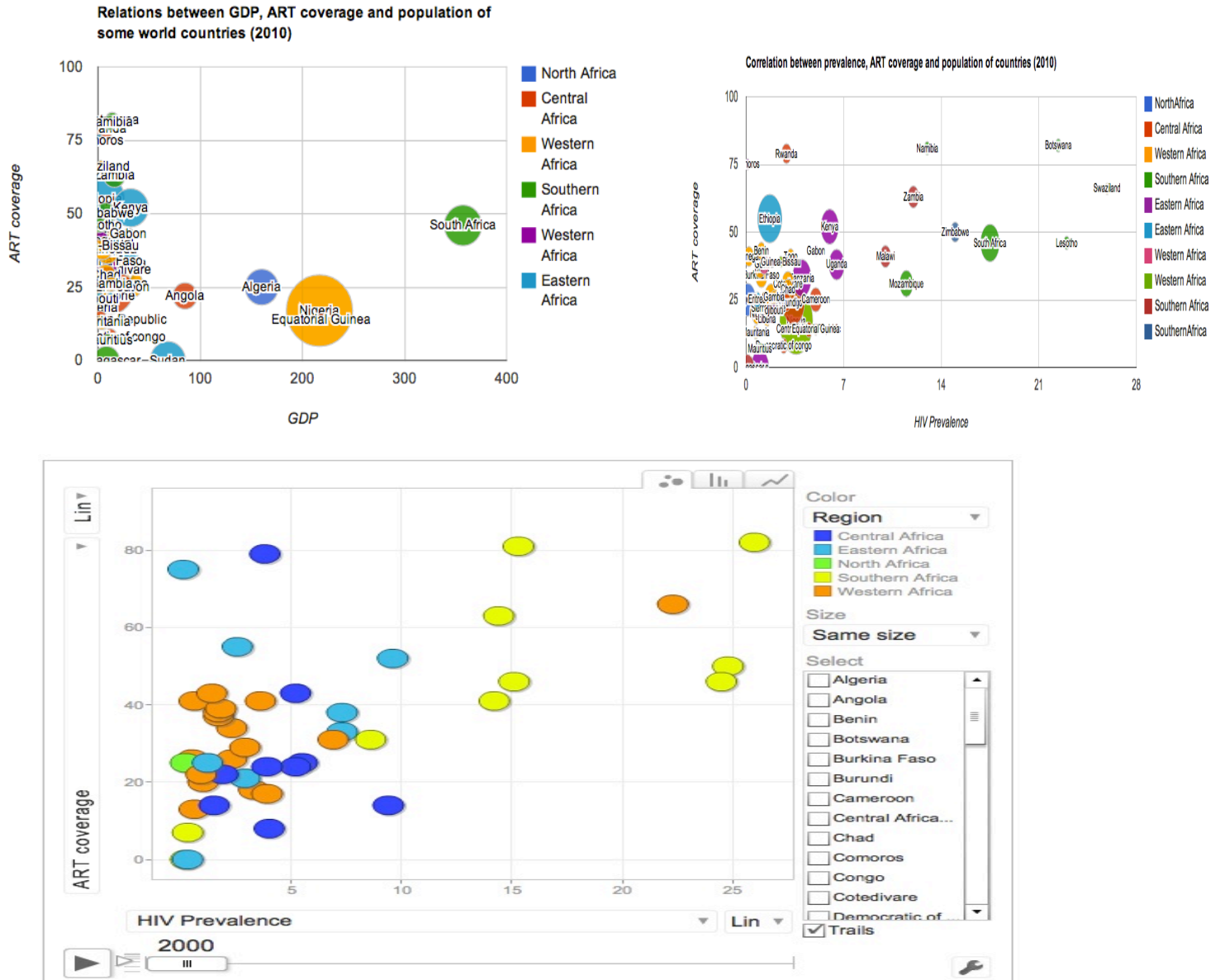


Figure 13: Indicator correlation visualization over LOHD system

5. Conclusion and lessons Learned

5.1. Summary

Linked Data is a new field of research to be used as a representation, visualization and intelligent searching method for complex heterogeneous data on the semantic web. In this thesis, we use the concepts of linked data and semantic web in the representation of the complex health related data in reusable and interoperable manner. The output shows that semantic web and LOD have a big potential in health data representation, visualization and querying. The proposed system was able to handle interoperability in syntactic and semantic level by the standardized and unique RDF vocabularies used in the data modeling.

Statistical data shaped in a form, which represents more than just numbers but also advanced meta-data with dimensions of the observation was represented using the standardized RDF data cube vocabulary.

There are different technologies to build linked data systems. RDF is a robust and flexible way for health data representation. Sgvizler, which supports all visualizations rendered by Google visualization tools is the best option for visualization. Sig.ma and other linked data tools are available for linked data query searching. In order to integrate those technologies and build a useful application for users, the four-layer approach, which support future scale up, proposed in this study can be used as a framework.

In conclusion, the set of technologies associated with semantics and ontologies over Linked Open Data in health care are, relatively speaking, still in their infancy. While there are high expectations, only modest progress has occurred to date. Nevertheless, from the output of this project, we can conclude that Linked data is the future potential technology for health data representation, querying and visualization but yet in order to get the best from it, improvements are needed both at the level of triple stores performance, domain-specific ontological vocabularies and visualization and searching tools performance.

5.2. Lessons learned and Future works

As linked data is a new science, there are obviously different challenges from modeling to system integration. In data conversion, I want to point out the complexity associated with the identification of groupings. In the case of multi-dimensional spreadsheets, information may be grouped using several distinct criteria, both in lines and columns, generating a very large of possible combinations. RDF vocabulary reuse is highly advisable, to promote interoperability and to help identify links to additional resources in the Linked Open Data cloud. This, of course, can only be achieved with adequate libraries and tool support to help identify possible matches to existing vocabularies. We worked with a few, domain specific vocabularies, but future plans include the incorporation of an ontology matching tool and the construction of an RDF vocabulary library. In line with that, our greatest challenge was the development of an algorithm that takes into consideration the structure, i.e., and the way data was organized in the spreadsheets, to enhance performance in tasks related to grouping identification and discovery.

Linked Data poses challenges inherent to querying highly heterogeneous and distributed data. To query linked data on the Web today, users must first be aware of which exposed datasets potentially contain the data they want and what data model describes these datasets, before using this information to create structured queries. This query paradigm is deeply attached to the traditional perspective of structured queries over databases and doesn't suit the linked data Web's heterogeneity, distributiveness, or scale. It's impractical to expect Web data consumers to have a previous understanding of available linked datasets structure and location that is still a challenge in linked data and is an open research question.

Future recommended works to improve the system includes; describing the data more robustly with domain specific additional vocabularies, improve the interlinking with more ontologies, improve the linked data search engine with better tools and add more visualization options for grouped data.

6. References

- [1] LIPPEVELD T (2001). ROUTINE HEALTH INFORMATION SYSTEMS: THE GLUE OF A UNIFIED HEALTH SYSTEM. KEYNOTE ADDRESS AT THE WORKSHOP ON ISSUES AND INNOVATION IN ROUTINE HEALTH INFORMATION IN DEVELOPING COUNTRIES, POTOMAC, MARCH 14–16.
- [2] MURRAY C, FRENK J (2000). A FRAMEWORK FOR ASSESSING THE PERFORMANCE OF HEALTH SYSTEMS. BULLETIN OF THE WORLD HEALTH ORGANIZATION, 79(6):717–732.
- [3] WHO (2000). HEALTH SYSTEM PERFORMANCE ASSESSMENT: REPORT BY THE SECRETARIAT. EB DOCUMENT 10/79.
- [4] ISSUES IN HEALTH INFORMATION (UNEDITED TEXT) NATIONAL AND SUBNATIONAL HEALTH INFORMATION SYSTEMS, BULLETIN OF THE WORLD HEALTH ORGANIZATION :6
- [5] BERNERS-LEE, T. (2006, JULY 27). LINKED DATA- DESIGN ISSUES. RETRIEVE FEBRUARY 2012, FROM [HTTP://WWW.W3.ORG/DESIGNISSUES/LINKEDDATA.HT ML](http://www.w3.org/DesignIssues/LinkedData.html)
- [6] JOCELYNE DO SACRAMENTO ,EKONG EMAH (MARCH, 2007), HIV/AIDS ASSESSMENT IN SUB-SAHARAN AFRICA DECUMENT. (2010).
- [7] SEMANTIC INTEROPERABILITY COMMUNITY OF PRACTICE: INTRODUCING SEMANTIC TECHNOLOGY AND VISION OF THE SEMANTIC WEB: MODULE 2
- [8] PETER J. GROEN, MARC WINE (JUNE 2009), MEDICAL SEMANTICS, ONTOLOGIES, OPEN SOLUTIONS AND EHR SYSTEMS, [HTTP://WWW.HOISE.COM/VMW/09/ARTICLES/VMW/LV-VM-09-09-6.HTML](http://www.hoise.com/vmw/09/articles/vmw/LV-VM-09-09-6.html)
- [9] GRUBER, T. (1993). A TRANSLATION APPROACH TO PORTABLE ONTOLOGY SPECIFICATIONS. KNOWLEDGE ACQUISITION (5), 199-220.
- [10] OASIS – WORKING DECUMENT- OPEN ARCHITECTURE FOR ACCESSIBLE SERVICES INTEGRATION AND STANDARDIZATION GRANT AGREEMENT # 215754 PAGE 35-40
- [11] J. BROEKSTRA, A. KAMPMAN, AND F. VAN HARMELEN. SESAME:A GENERIC ARCHITECTURE FOR STORING AND QUERYING RDF AND RDFSHEMA. IN FIRST INTERNATIONAL SEMANTIC WEB CONFERENCE (ISWC 2002), PAGES 54–68. SPRINGER BERLIN / HEIDELBERG, 2002. 10.1007/3-540-48005-6_7.
- [12] CHRISTIAN BIZER AND ANDREAS SCHULTZ. THE R2R FRAMEWORK: PUBLISHING AND DISCOVERING MAPPINGS ON THE WEB. IN PROCEEDINGS OF THE 1ST INTERNATIONAL WORKSHOP ON CONSUMING LINKED DATA, 2010. 25, 102

-
- [13] NUREFŞAN GÜR , LAURA DIAZ SANCHEZ , TOMI KAUPPINEN GI SYSTEMS FOR PUBLIC HEALTH WITH AN ONTOLOGY BASED APPROACH: PROCEEDINGS OF THEAGILE'2012 INTERNATIONAL CONFERENCE ON GEOGRAPHIC INFORMATION SCIENCE, AVIGNON, APRIL, 24-27, 2012
- [14] AMRAPALI ZAVERI, RICARDO PIETROBON,SÖREN AUER ,JENS LEHMANN,MICHAEL MARTIN TIMOFEY ERMILOV: ReDD-OBSERVATORY: USING THE WEB OF DATA FOR EVALUATING THE RESEARCH-DISEASE DISPARITY FOR EMERGING REGIONS
- [15] AMRAPALI ZAVERI, RICARDO PIETROBON,SÖREN AUER ,JENS LEHMANN,MICHAEL MARTIN TIMOFEY ERMILOV: PUBLISHING AND INTERLINKING THE GLOBAL HEALTH OBSERVATORY:SEMANTIC WEB 1(2012) 1-5.
- [16] RAIMOND, Y., SUTTON, C., SANDLER, M.: AUTOMATIC INTERLINKING OF MUSIC DATASETS ON THE SEMANTIC WEB. IN: LINKED DATA ONTHE WEB WORKSHOP (LDOW2008), 2008.
- [17] HASSANZADEH, O., ET AL.: A DECLARATIVE FRAMEWORK FOR SEMANTIC LINK DISCOVERY OVER RELATIONAL DATA. POSTER AT 18TH WORLD WIDE WEB CONFERENCE (WWW2009), 2009.
- [18] LU, J. ET AL. LEARNING DEEP WEB CRAWLING WITH DIVERSE FEATURES. IN: PROC. 2009 IEEE/WIC/ACM INTERNATIONAL JOINT CONFERENCE ON WEB INTELLIGENCE AND INTELLIGENT AGENT TECHNOLOGY - VOLUME 01, WI-IAT '09 (2009): 572-575.
- [19] CASANOVA, M.A., BREITMAN, K., BRAUNER, D.F. AND MARINS, A. DATABASE CONCEPTUAL SCHEMA MATCHING. COMPUTER (LONG BEACH), v. 40 (2007): 102-104.
- [20] BELL, G., HEY, T. AND SLAZAY, A. BEYOND THE DATA DELUGE SCIENCE (MARCH 2009): 1297-1298
- [21] HEY, TANSLEY, TOLLE (EDS) THE FOURTH PARADIGM - MICROSOFT RESEARCH (2009).
- [22] BREITMAN, K., CASANOVA, M.A., AND TRUSZKOWSKI, W. SEMANTIC WEB: CONCEPTS, TECHNOLOGIES AND APPLICATIONS.LONDON: SPRINGER, 2006. v. 1. 337 p. 28
- [23] KINSELLA, S., BOJARS, U., HARTH, A., BRESLIN, J.G. AND DECKER, S. AN INTERACTIVE MAP OF SEMANTIC WEB ONTOLOGY USAGE. IN: INFORMATION VISUALISATION, 2008. IV '08. 12TH INTERNATIONAL CONFERENCE (9-11 JULY 2008): 179-184.

-
- [24] AUER, S., DIETZOLD, S., LEHMANN, J., HELLMANN, S. AND AUMUELLER, D. TRIPLIFY: LIGHTWEIGHT LINKED DATA PUBLICATION FROM RELATIONAL DATABASES. IN: PROC. 18TH INTERNATIONAL CONFERENCE ON WORLD WIDE WEB. MADRID, SPAIN: ACM. (2009): 621-630.
 - [25] ERLING, O. AND MIKHAILOV, I. RDF SUPPORT IN THE VIRTUOSO DBMS. IN: PROC. 1ST CONFERENCE ON SOCIAL SEMANTIC WEB, VOLUME P-113 OF GI-EDITION - LECTURE NOTES IN INFORMATICS (LNI), BONNER KOLLEN VERLAG (SEPT. 2007).
 - [26] BIZER. C. AND SEABORNE, A. D2RQ - TREATING NON-RDF DATABASES AS VIRTUAL RDF GRAPHS. IN ISWC2004 (POSTERS), (NOV. 2004).
 - [27] GOULD P. THE SLOW PLAGUE: A GEOGRAPHY OF THE AIDS PANDEMIC. OXFORD, UNITED KINGDOM AND CAMBRIDGE, USA: BLACKWELL PUBLISHERS, 1993.
 - [28] SHANNON GW. THE GEOGRAPHY OF AIDS: ORIGINS AND COURSE OF AN EPIDEMIC. NEW YORK: GUILFORD PRESS, 1991.
 - [29] WEIR SS, MORRONI C, COETZEE N, SPENCER J, BOERMA JT. A PILOT STUDY OF A RAPID ASSESSMENT METHOD TO IDENTIFY PLACES FOR AIDS PREVENTION IN CAPE TOWN, SOUTH AFRICA. SEX TRANSM INFECT. 2002;78:106-113.
 - [30] CARREL M, ESCAMILLA V, MESSINA J, GIEBULTOWICZ S, WINSTON J, YUNUS M, STREATFIELD PK, EMCH M. DIARRHEAL DISEASE RISK IN RURAL BANGLADESH DECREASES AS TUBEWELL DENSITY INCREASES: A ZERO-INFLATED AND GEOGRAPHICALLY WEIGHTED ANALYSIS. INT J HEALTH GEO. 2011;6:10-41.
 - [31] BROOKER, S. SPATIAL EPIDEMIOLOGY OF HUMAN SCHISTOSOMIASIS IN AFRICA: RISK MODELS, TRANSMISSION DYNAMICS AND CONTROL. TRANS ROYAL SOCIETY TROP MED HYGIENE. 2007; 101 (1):1-8.
 - [32] KEINSCHMIDT I, PETTIFOR A, MORRIS N, MACPHAIL C, REES H. GEOGRAPHIC DISTRIBUTION OF HUMAN IMMUNODEFICIENCY VIRUS IN SOUTH AFRICA. AM J TROP MED HYG. 2007;77(6):1163-1169.
 - [33] KALIPENI E, ZULU L. USING GIS TO MODEL AND FORECAST HIV/AIDS RATES IN AFRICA, 1986-2010. PROF GEOG. 2008;60 (1):33-53.
 - [34] AJJAMPUR SSR, GLADSTONE BP, SELVAPANDIAN D, MULIYIL JP, WARD H, KANG G. MOLECULAR AND SPATIAL EPIDEMIOLOGY OF CRYPTOSPORIDIOSIS IN CHILDREN IN A SEMIURBAN COMMUNITY IN SOUTH INDIA. J CLIN MICROB. 2007;45 (3):915-920.
 - [35] MARCIA LUCAS PESCE, KARIN KOOGAN BREITMAN, MARCO ANTONIO CASANOVA: SURFACING SCIENTIFIC AND FINANCIAL DATA WITH THE XCEL2RDF PLUG-IN : PAGE 4-12

-
- [36] MAPPING SEMANTIC WEB DATA WITH RDBMSes
<[HTTP://WWW.W3.ORG/2001/SW/EUROPE/REPORTS/SCALABLE_RDBMS_MAPPING_REPORT/](http://www.w3.org/2001/sw/europe/reports/scalable_rdbms_mapping_report/)>
- [37] ANJA JENTZSCH, ROBERT ISELE, CHRISTIAN BIZER (2009) GENERATING RDF LINKS WHILE PUBLISHING ORCONSUMING LINKED DATA: 2009
- [38] J. VOLZ, C. BIZER, M. GAEDKE, AND G. KOBILAROV. DISCOVERING AND MAINTAINING LINKS ON THE WEB OF DATA. IN INTERNATIONAL SEMANTIC WEB CONFERENCE, PAGES 650{665, 2009.
- [39] OLAF HARTIG AND ANDREAS LANGEgger. A DATABASE PERSPECTIVE ON CONSUMING LINKED DATA ON THE WEB. DATENBANK-SPEKTRUM, 10:57–66, 2010. 10.1007/s13222-010-0021-7. [HTTP://DX.DOI.ORG/10.1007/S13222-010-0021-7](http://dx.doi.org/10.1007/s13222-010-0021-7)DOI: 10.1007/s13222-010-0021-7
- [40] OLAF HARTIG, HANNES MUEHLEISEN, AND JOHANN-CHRISTOPH FREYTAG. LINKED DATA FOR BUILDING A MAP OF RESEARCHERS. IN PROCEEDINGS OF THE 5TH WORKSHOP ON SCRIPTING FOR THE SEMANTIC WEB, 2009.
- [41] KHROUF, H., ATEMEZING, G., RIZZO, G., TRONCY, R., & STEINER, T. (2012). AGGREGATING SOCIAL MEDIA FOR ENHANCING CONFERENCE EXPERIENCE. (ICWSM'12) 1ST INTERNATIONAL WORKSHOP ON REAL-TIME ANALYSIS AND MINING OF SOCIAL STREAMS (RAMSS'12). DUBLIN, IRELAND, JUNE 4, 2012. LIEBERMAN, J. (2010, JULY 21).
- [42] A. CHEPTSOV, M. ASSEL, G. GALLIZO, I. CELINO, D. DELLAGLIO, L. BRADSKO, M. WITBROCK, AND E. DELLA VALLE. LARGE KNOWLEDGE COLLIDER. A SERVICE-ORIENTED PLATFORM FOR LARGE-SCALE SEMANTIC REASONING. IN PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON WEB INTELLIGENCE, MINING AND SEMANTICS (WIMS2011), 2011.
- [43] MARTIN G. SKJ_VELAND SGVIZLER: A JAVASCRIPT WRAPPER FOR EASY VISUALIZATION OF SPARQL RESULT SETS: SEMANTIC WEB JOURNAL(2010)
- [44] GIOVANNI TUMMARELLO, RICHARD CYGANIAK, MICHELE CATASTA, SZYMON DANIELCZYK, RENAUD DELBRU, AND STEFAN DECKER. SIG.MA: LIVE VIEWS ON THE WEB OF DATA. WEB SEMANTICS: SCIENCE, SERVICES AND AGENTS ON THE WORLD WIDE WEB, 8(4):355 – 364, 2010. [HTTP://DX.DOI.ORG/10.1145/1772690.1772907](http://dx.doi.org/10.1145/1772690.1772907)DOI:10.1145/1772690.1772907