Potential of Linked Open Data in Health Information Representation on the Semantic Web

Binyam Tilahun

Institut für Geoinformatik, Universität Münster Wesselerstraße. 253,
Münster, Germany
binyam.tilahun @uni-muenster.de

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 is a new semantic web standard to publish and link heterogonous data. Its representation allows not only human, but also machine to browse the data in unlimited way. Through a use case of world health organization data, this project shows that current linked open data technologies have a potential to represent heterogonous health data in a flexible 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 health data representation.

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

1. Introduction

The foundation of effective decision-making is the effective use of data [1]. 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.

The World Wide Web, which is a system of interlinked documents[1], has changed the way we share information among people. It 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.

1 http://www.w3.org/WhatIs.html accessed on October 3,2012

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 health official data databases, exploiting the different kinds of available information about a given topic is challenging because data are spread over the World Wide Web (Web), in different databases, hosted in a large number of independent, heterogeneous and highly focused resources. In our case, when you look for specific data about HIV, there are different official sources like WHO, CDC or UNAIDS with different level of representation and interests. Those challenges in health information data access require new harmonized, integrated and interoperable way of health data representation and dissemination.

The Semantic Web provides a common framework that allows data to be shared and re-used across application, enterprise, and community boundaries for different purpose and interest. The Semantic Web is an evolving extension of the World Wide Web (WWW) in which web content can be 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 [2]. At its core, the Semantic Web comprises a philosophy, a set of design principles, collaborative working groups, and a variety of enabling technologies. Some elements of the semantic web are expressed in formal specifications such as the Resource Description Framework (RDF), a variety of data interchange formats (e.g. RDF/XML, N3, Turtle), and notations such as RDF Schema and the Ontology Web Language (OWL), all of which are implicitly or explicitly used to provide a formal description of concepts, terms, and relationships within a given knowledge domain.

The Resource Description Framework (RDF) model [3] is based upon the notion of making descriptive statements about resources. A RDF statement, also called a triple in RDF terminology is an association of the form *(subject, predicate, object)*. Subject and object are both URIs describing an object as well predicate is also a URI defining the relations between subject and object [4]. The predicate is a resource as well, denoting a specific property of the subject. The object, which can be a resource or a string literal, represents the value of this property.

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.

2. Dataset preparation and conversion

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 data set in its global health observatory data repository[1], and missing data for some years was complemented by 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. The goal is to build linked HIV data allowing users to query and visualize about the current trend of HIV in HIV-burdened region of the world, sub Saharan Africa.

The statistical data about each indicator for each country (like prevalence, incidence, mortality and data on the response) was converted into RDF using the LOD concept of subject, predicate and object. To make the data richer, it was linked to existing data by using existing ontological vocabularies of LinkedCT, MeSH Vocabulary, DBPedia, GHO, PubMed and Diseasome networks. Linking the data to these existing vocabularies using the predicate owl:sameAs enriches the local data. By the potential of this predicate two data which doesn't know each other with a link reference will implicitly known by the rest of the data linked to the subject which in turn will make the data more rich. Silk, a link discovery framework for a web of data was applied to get important relationships between the data. Silk provides a declarative language for specifying the types of RDF links that should be discovered between data sources and the conditions which the data items must fulfill in order to be interlinked[2]. Jaro distance was used as string metric in the link discovery process.

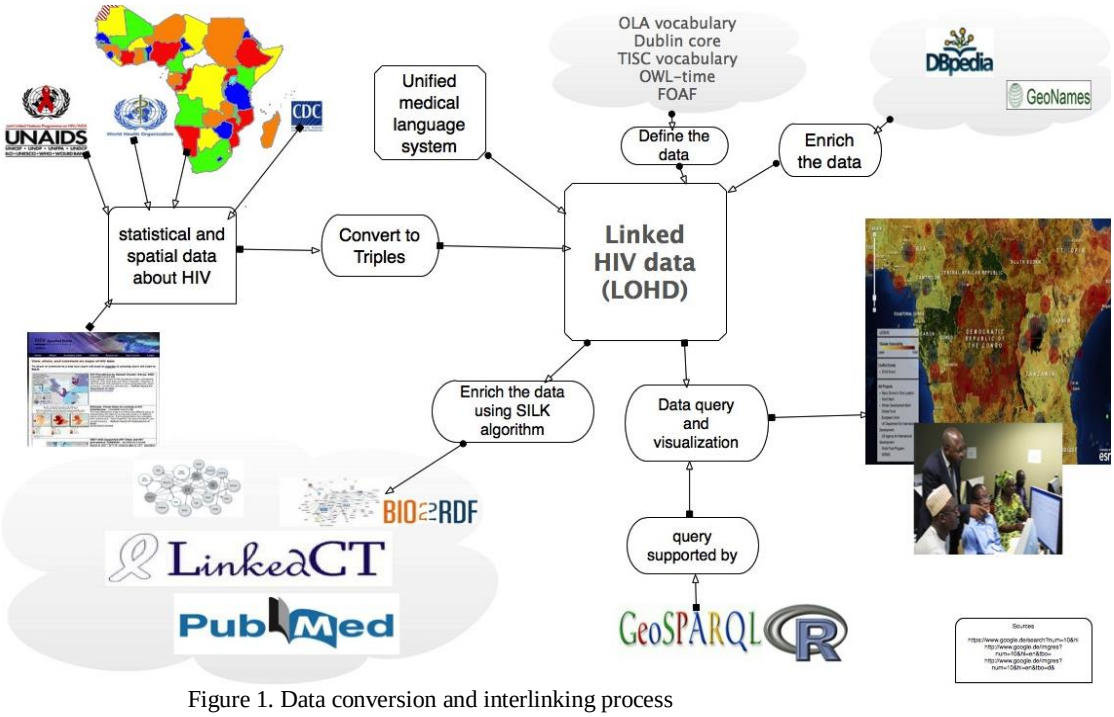

Figure 1. Data conversion and interlinking process

[1] http://apps.who.int/ghodata/?vid=22100 Accessed on October 5,2012

[2] http://wifo5-03.informatik.uni-mannheim.de/bizer/silk/

3. Information retrieval with SPARQL

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, timeline and visualizations. Triples stored in the parliament database, 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. Current search engine of Linked Data is 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.

4. Conclusion and further works

The main goal of the work was to try to apply the concepts of linked data and semantic web in representation of the complex health related data in reusable and interoperable manner. The output shows that semantic web and LOD had a potential in health data representation, visualization and querying. The set of technologies associated with semantics and ontologies in health care are, relatively speaking, still in their infancy. While there are high expectations, only modest progress has occurred to date. Further works will be, making the data more rich and implement better querying and visualization tools, which will make it easy to use by health professionals and decision makers.

5. References

1. *AbouZahr C, Boerma T. Health information systems: the foundations of public health.* Bull World Health Organ *2005; 83: 578-83 pmid: 16184276.*
2. *T. Berners-Lee and J. Hendler, Publishing on the semantic web, Nature, vol. 410, pp. 1023-1024, 2001.*
3. *S.R. Bratt, Toward a Web of Data and Programs, in IEEE Symposium on Global Data Interoperability - Challenges and Technologies, 2005.*
4. *J. Davies, Semantic Web Technologies: Trends and Research in Ontology-based Systems: John Wiley & Sons, 2006.*
5. *C. Bizer, T. Heath, T. Berners-Lee, Linked Data - The Story So Far, Int .J. on Semantic Web and Information Systems 5 (3) (2009)*