Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 164 (2019) 686–691

Procedia Computer Science

www.elsevier.com/locate/procedia

CENTERIS - International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies

Integrating Tuberculosis data in State of São Paulo over Semantic Web: a proof of concept

Felipe Carvalho Pellison[a,*], Vinícius Costa Lima[a], Rui Pedro Chartes Lopes Rijo[b,c,d], Domingos Alves[c,e]

[a]*Bioengineering Postgraduate Program, University of São Paulo, São Carlos, Brazil*
[b]*School of Technology and Management, Polytechnic Institute of Leiria, Leiria, Portugal*
[c]*Center for Research in Health Technologies and Information Systems (CINTESIS), University of Porto, Porto, Portugal*
[d] *Institute for Systems and Computers Engineering at Coimbra (INESCC), Coimbra, Portugal*
[e] *Ribeirão Preto Medical School, University of São Paulo, Ribeirão Preto, Brazil*

Abstract

Although tuberculosis is a curable disease and, in most cases, with low cost drugs, its mortality still is a global concern. This facts turns our attention to management issues and the difficulties related to retrieving data of interest that are powdered on many applications. This work presents a semantic web approach to achieve functional and semantic interoperability between two applications in State of São Paulo that contain tuberculosis data. By combining a theoretical-practical development, the geolocalization tool created is a proof of concept that could help managers to take strategic decisions and develop better health policies by showing the distribution of tuberculosis cases across the state. This work stands out the importance of working in solutions that could improve the quality of data in health field and daily activities of health professionals.

[*] Corresponding author.
E-mail address: felipecp@alumni.usp.br

1877-0509 © 2019 The Authors. Published by Elsevier B.V.

10.1016/j.procs.2019.12.236

Keywords: Tuberculosis; Semantic Web; Health Information System; Interoperability; Data Integration

1. Introduction

One of the key issues regarding Health Information Systems is its lack of interoperability. Systems do not exchange data, and even when they do so, their data has not the same meaning. This situation leads to disconnected services operations, re-work and poor quality of the health services. All these factors limit the possibility of using data for scientific studies in the area.

As a world health concern matter, we have Tuberculosis (TB). TB is a curable disease, but in 2016, it was among the top 10 causes of death, with 1.3 million deaths and about 10.4 million new cases worldwide [1]. In 2016, Brazil notified 75 thousand new cases and in 2014, registered a cure date of 72.8%, 10.5% of abandonment and 7.8% of death [2].

In State of São Paulo, in Brazil, tuberculosis deals with, at least, five main health information systems that run via web, namely, SISTB, Hygia Web, TBWeb, SINAN, GAL and e-SUS AB. In some of the applications, health professionals have to re-introduce the same information. When global information about a patient is needed, some specific ad hoc search work has to be done in each one of these applications. All this scenario leads to many difficulties concerning the management, the consistency and the integrity of the data that is handled.

The data distributed among all these applications can be used to draw epidemiological studies that can serve as a basis for new strategies and policies on every level of healthcare. For instance, visualizing the location of every notification of TB in a fast way can help health managers to take fast actions to distribute the right amount of resources for specific locations. Indeed, retrieving the data of interest may me quite a challenge, because interoperability is not available for these applications.

Thus, this work aims to present an approach to integrate TB related data in State of São Paulo, Brazil, gathered from heterogeneous data sources through Semantic Web paradigm, due to its flexible governance, low cost of the technicals changes that was necessary to implement on each of the health applications to provide the data interoperability and the possibility to make semantically marked up health data readable by machines. A proof of concept will be developed to underpin the importance of combining and consolidating health information to enhance data quality and daily basis processes in health services, seeking better outcomes in TB care.

In the next section, key resources and concepts will be introduced. In the third section, the research methods of this study will be described. In the fourth section, results will be showed, followed by a discussion. Finally, in the last section, final considerations will be presented, including possible limitations and future work.

2. Background

2.1. Semantic Web

In today's web, most of the available content is suitable for human interpretation and is therefore not easily accessible by other machines and systems. The semantic web, defined by Tim Berners-Lee in [3], can be specified as an extension of the current web, with the purpose of adding logic to the content to express the meaning of an information, its properties and the complex relationships existing between different types of data, so that it is possible to interpret the meaning of a given data without worrying about its form of representation [4]. The general idea is to create an efficient way to represent data on the World Wide Web to build a global database of connected data [5], through the semantic marking of web pages and the adaptation of existing relational databases, based on ontologies.

The basic blocks that define the semantic web are a standard data model, a query protocol, and a set of reference vocabularies. W3C standards and definitions, such as the Resource Description Framework (RDF), Simple Protocol and RDF Query Language (SPARQL), and ontologies refer to these basic blocks, defined as a description language and data model, a query protocol to obtain data stored in RDF and a formal representation of a given knowledge, respectively in [6]. Ontologies can be defined as a formal representation of knowledge in a specific domain [7], aiming to formulate a rigorous and exhaustive conceptual scheme. In turn, Web Ontology Language (OWL) is a semantic markup language for publishing and sharing ontologies, designed to describe classes and relationships between them [8].

In order to mark up data and serialize them, we used JSON-Linked Data (https://www.w3.org/TR/2014/REC-json-ld-20140116/). It was a logical choice because he is easily integrated with systems that already uses JSON, enabling machines to interpret data by describing contexts and properties.

2.2. Tuberculosis Information Systems

SisTB is a regional web-based system for monitoring TB patients in the city area of Ribeirão Preto, State of São Paulo. This system allows patients registration and treatment follow-up through the Directly Observed Treatment, Short Course (DOTS) strategy, a highly efficient and cost-effective strategy recommended by the World Health Organization (WHO) [9]. Additionally, SisTB is part of a set of integrated systems that aims to deliver a better management of TB cases. SisTB relies on a functional and semantic interoperability architecture based on standards preconized by the World Wide Web Consortium (W3C), such as the Semantic Web, allowing optimized data extraction by machines through APIs and semantic query endpoints [10].

TBWeb, the Tuberculosis Notification and Follow-Up System, is a web-based system for TB cases notification, basic patient follow-up (does not implement DOTS strategy) and epidemiologic vigilance. All patients diagnosed with TB in State of São Paulo are registered in this system.

3. Research Methods

Action Research will be the scientific methodological approach for this study. The investigative and practical procedures proposed by this methodology were considered the most appropriate, considering that the project has a expected theoretical-practical development. By choosing this methodology, it is expected to improve an observed context, i.e., TB data integration through Semantic Web, and simultaneously expand scientific knowledge through practical problem-solving, enhancing relevant competencies of involved actors [11].

3.1. Overview

The novelty of this work is to present a proof of concept regarding TB data integration in the State of São Paulo, Brazil, using Semantic Web resources, such as SPARQL queries and RDF. Throughout a federated query, data will be simultaneously obtained from TBWeb, the governmental system, and from SisTB, the regional TB information system used mainly in city of Ribeirão Preto. By doing this, it will be possible to combine data from both sources with aggregated semantic value. Also, divergences among pieces of data will be highlighted and users will be able to explore this data visually.

Hopefully, this proof of concept will clarify the potential of Semantic Web applied to health information systems to integrate data from heterogeneous data sources, aiming to expand this approach at a national level to other TB relevant information systems.

3.2. Technological tools

In order to achieve the expected results, a set of tools, protocols and programming languages were used. All selected resources are non-proprietary.

In the frontend development, that basically represent the graphical interface, HTML5, JavaScript and JSTL (JavaServer Pages Standard Tag Library) Expression Language were used to build the web page. Bootstrap framework components were applied to make it responsible. Finally, Bing Maps API (https://www.bingmapsportal.com/) was chosen to delivery the map functionality.

In the backend development, Java programming language, Enterprise Edition Version 8, was used, since there are frameworks available to boost semantic applications, such as Apache Jena (https://jena.apache.org/), which supply interfaces, classes and methods necessary to work with semantic data in RDF and others semantic formats.

SisTB interoperability architecture include a SPARQL endpoint, powered by D2RQ Server (http://d2rq.org/). The D2RQ software allows accessing relational databases as virtual RDF through SPARQL queries. Furthermore, the SPARQL endpoint can be used to perform federated queries in distinct databases, which is crucial to semantic data integration.

3.3. Data Sample

For this work development, a anonymized data sample from TBWeb of 7277 TB cases notified in State of São Paulo, from 2006 to 2014, was used, along with the whole SisTB dataset. A common identifier was chosen to map TB patients these datasets. This identifier is called SINAN ID, that is a number generated by a national information system for patients with diseases of compulsory notification.

Demographic data were used to push points on a map and to compare values among datasets, which include: latitude, longitude, pregnancy situation, age, gender, city of notification, federative unity (state), scholarity and ethnicity.

Finally, to select data across the systems, a simple ontology was used, designed specifically to address this data and for testing purposes. The ontology is used through D2RQ Server, which maps the relational database schema and content. Thus, the server exposes data as a virtual RDF graph that can be queried using SPARQL protocol.

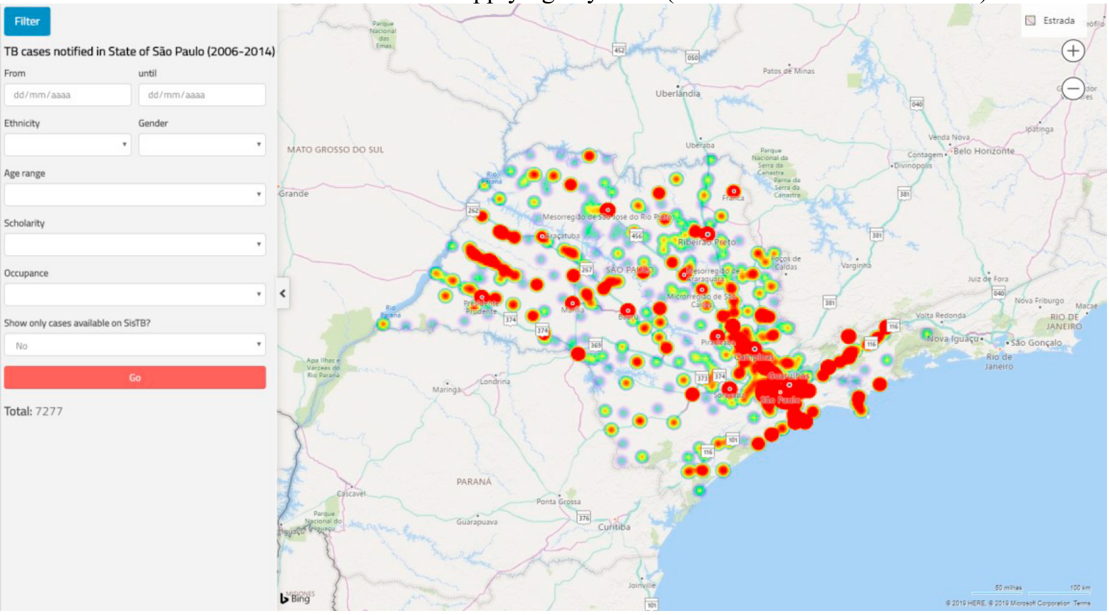
The ontology was based on schema.org Person ontology (http://schema.org/Person), but considering that the design was not in this study scope, its development will not be represented.

4. Results and Discussion

By applying the presented methodology, an interactive tool with a graphical user interface was developed to a better visualization of TB data in State of São Paulo, Brazil. This tool will be able to gather data from two health information systems, SisTB and TBWeb, through Semantic Web paradigm.

The user is able to obtain information about notification of TB cases using filters available in the interface. The search result will be drawn in a map as a heatmap, according to the municipalities that have notified TB cases. One special filter, named as "Show only cases available on SisTB?", can be used to indicate that data should be combined in both target systems. By doing this, only cases that are simultaneously in both systems will be retrieved.

Fig. 1 shows a search execution result without applying any filter (all data available on TBWeb).


Filter

TB cases notified in State of São Paulo (2006-2014)

From: dd/mm/aaaa   until: dd/mm/aaaa

Ethnicity: ▾   Gender: ▾

Age range: ▾

Scholarity: ▾

Occupance: ▾

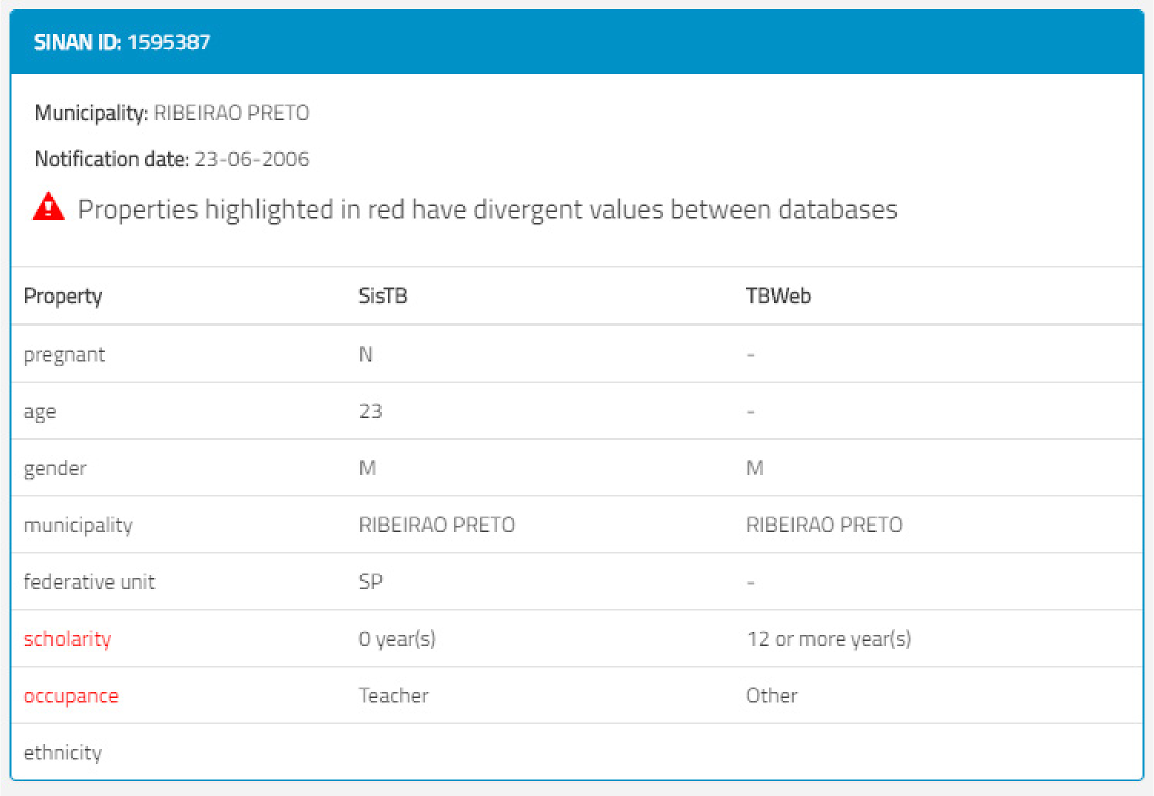
Show only cases available on SisTB?: No ▾

Go

Total: 7277

Fig. 1 - Example of a search execution

If the user does not trigger a search over SisTB and TBWeb, a SPARQL query is generated to get only data from TBWeb to build the heatmap. However, when the user triggers a combined search, after finishing execution, a new option "Explore data" appears, which allows navigation through the cases notified in TBWeb and followed in SisTB. Data is obtained by sending a federated SPARQL query and is presented in a table. By clicking in SINAN ID (first column), the information stored in each database will be showed in a new table for comparison and discrepancies will be highlighted, as shown in Fig. 2. In this computational experiment, it was possible to identify 126 cases notified in TBWeb that were followed in SisTB. The interoperability architecture available on SisTB is responsible for executing the SPARQL queries.



SINAN ID: 1595387

Municipality: RIBEIRAO PRETO

Notification date: 23-06-2006

⚠ Properties highlighted in red have divergent values between databases

Property	SisTB	TBWeb
pregnant	N	-
age	23	-
gender	M	M
municipality	RIBEIRAO PRETO	RIBEIRAO PRETO
federative unit	SP	-
scholarity	0 year(s)	12 or more year(s)
occupance	Teacher	Other
ethnicity		

Fig. 2 - Comparison of data retrieved through the federated SPARQL query

By identifying divergence and errors in information stored in both databases, data quality could be enhanced through manual manipulation or using an automatized procedure to perform updates based on rules. Furthermore, in an integrated environment, rework of health professionals could be reduced by eliminating the need of inputting the same data in several systems. Consolidated data among health information systems could lead to a valuable basis to the development of new clinical and managerial decision-support tools.

The Semantic Web plays an important role in assimilating data from heterogeneous data sources due to its resources, such as ontologies, to aggregate meaning to data, SPARQL protocol, that allows querying over these data sources, and RDF graphs, which facilitates the modelling and transmission of semantic data.

5. Final Considerations

This proof of concept calls attention to the importance of working in solutions that could improve the quality of data in health field and daily activities of health professionals. Combining several datasets reveals to be an interesting approach to achieve that and the Semantic Web represents a key component due to its available resources. The initiative proposed in this work could be expanded to a national level, where every system and database of TB data could be part of an integrated environment.

As a possible limitation of this study, obtaining access authorization to systems and databases can be a bureaucratic barrier, considering that it involves negotiation of security aspects and possible costs for customization, mainly due to lack of interoperability features. However, discussion are being carried out with the Health Department of the State of São Paulo and the Brazilian Ministry of Health to define relevant entities and to negotiate technical aspects and human resources to a next phase, which will consist in the validation and expansion of the presented methodology.

Finally, as future work, it is intended to develop appropriate ontologies to a better data representation and to effectively address concepts over all systems that could be integrated, as well as defining requirements to build useful tools based on the needs of managers and health professionals.

Acknowledgements

This research is under development in the Bioengineering Postgraduate Program (EESC/FMRP/IQSC) of University of São Paulo. DA would like to thank the São Paulo Research Foundation (FAPESP) for financial support (Process number 2018/00307-2)

References

[1] WHO, Global Tuberculosis Report, 2017.

[2] WHO, "WHO Global tuberculosis report, 2016.

[3] T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web. A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities," Sci. Am., vol. 284, no. 5, pp. 34–43, 2001

[4] I. Robu, V. Robu, and B. Thirion, "An introduction to the Semantic Web for health sciences librarians.," J. Med. Libr. Assoc., vol. 94, no. 2, pp. 198–205, 2006.

[5] C. Laufer, "Guia_Web_Semantica," p. 133, 2015.

[6] W3C, "SPARQL Query Language for RDF," 2003. Available: https://www.w3.org/TR/rdf-sparql-query/.

[7] E. Craig, Routledge Encyclopedia of Philosophy. New York: Routledge, 1998.

[8] W3C, "OWL - Semantic Web Standards."Available: https://www.w3.org/OWL

[9] World Health Organization, What is DOTS ? A Guide to Understanding the WHO-recommended TB Control Strategy Known as DOTS, Prev. Control. (1999) 1–39.

[10] F.C. Pellison, R.P.C. Lopes Rijo, V.C. Lima, R.R. De Lima, R. Martinho, R.J. Cruz Correia, D. Alves, Development and evaluation of an interoperable system based on the semantic web to enhance the management of patients' tuberculosis data, Procedia Comput. Sci. 121 (2017) 791–796. doi:10.1016/j.procs.2017.11.102.

[11] M. Hult, S. Lennung, Towards a Definition of Action Research: a Note and Bibliography, J. Manag. Stud. 17 (1980) 241–250. doi:10.1111/j.1467-6486.1980.tb00087.x.