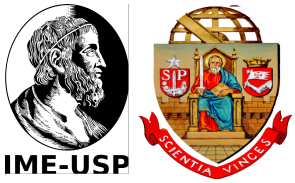

University of São Paulo - Brazil
The Institute of Mathematics and Statistics
Department of Computer Science

Semantic Data Integration for Public Health in Brazil*

Debora Lina Ciriaco dciriaco@ime.usp.br
Alexandre Pessoa alexcpp@ime.usp.br
Laís Salvador laisns@dcc.ufba.br
Renata Wassermann renata@ime.usp.br

ABSTRACT

The lack of semantic information is a big challenge, even in context-driven areas like Healthcare, characterized by established terminologies. Here, semantic data integration is the solution to provide precise information and answers to questions like: What is the care pathway of newborns diagnosed with a congenital anomaly in consequence of congenital syphilis in the city of São Paulo? This project will use a semantic data integration technique, ontology based data integration, to integrate three health databases from the city of São Paulo - Brazil: mortality, live births and hospital information system. It is expected that the integration of public health databases will help to map patient care pathways, predict public resource needs and minimize unnecessary spending.

Keywords: Semantic Data Integration, Ontology Based Data Integration, Ontologies, Healthcare, Public Health.

1 INTRODUCTION

The Brazilian Ministry of Health, through DATASUS (Informatics Brazilian Health System Department), requires more than 45 systems for reporting the national health situation[1]. Among them, there are systems for monitoring the demographic situation, such as the SIM (Mortality Information System)[2] and the SINASC (Information System on Live Births)[3]. Moreover, there are other systems for the collection of medical procedures of SUS (Brazilian Health System), such as the SIH (Hospital Information System)[4]. It can be noted that a large number of these systems do not necessarily interact with each other, guiding to independent, redundant and non-interoperable information [1].

Regarding health information management and decision making, one of the main difficulties is the access to these data sets, specifically making a single query in more than one database and performing complex queries that require distributed data. In this way, it is extremely costly to follow patients over time. Thus, it is not easily known how many patients are being treated by the health system nor what procedures are repeated by the same individual. While national initiatives have been undertaken [2–4] (such as the recent updates of information systems by the Ministry of Health), much work remains to be done. Most of the projects ignored semantic meaning [5–7] or were made on small cities [8].

*This research is part of the INCT of the Future Internet for Smart Cities funded by CNPq proc. 465446/2014-0, CAPES proc. 88887.136422/2017-00, and FAPESP proc. 14/50937-1 and FAPESP proc. 15/24485-9.

[1]DATASUS http://datasus.saude.gov.br/sistemas-e-aplicativos

[2]SIM: http://www2.datasus.gov.br/DATASUS/index.php?area=060701

[3]SINASC: http://www2.datasus.gov.br/DATASUS/index.php?area=060702

[4]SIH: http://datasus.saude.gov.br/sistemas-e-aplicativos/hospitalares/sihsus

2 PROPOSAL

Given the scenario and the open topics about the theme, this project aims to develop an ontology layer to integrate distinct health-care databases and, consequently, to do complex analysis, such as to track the patient care pathway (yet ensuring anonymity), allowing better resources' management. Here we will use a hybrid OBDI approach [9], developed by [10], to integrate SINASC, SIM and SIH databases and answer competence questions like:

- What is the level of education of the mother when the child was born?
- Did the mother have any records of hospitalization during pregnancy? Is hospitalization related to the pregnancy?
- Did the mother have any records of ICU admission?
- What is the care pathway of newborns diagnosed with a congenital anomaly in consequence of congenital syphilis in the city of São Paulo?

3 SEMANTIC DATA INTEGRATION

Different approaches can be used for integrating data, such as schema mapping and matching, model management, answering queries using views, data exchange, record linkage, data fusion, etc. [11]. However, the use of a semantic-based approach is particularly beneficial for data from complex urban environments. The data repositories across government agencies use different data models and schemas. In this scenario, it is common to encounter syntactic and semantic discrepancies, mainly due to spatial, temporal, and/or thematic diversities in the studied datasets. Ontology Based Data Integration - OBDI, is useful for solving the absence of interoperability between the databases because they are able to identify and associate the semantic correspondence of the concepts [12]. This kind of solution provides a coherent representation framework that simplifies data usage and unlocks their combined value [13]. The integration resolves heterogeneities with respect to the schemas and their data, either to enable their direct manipulation or to enable the automatic translation of data and queries across the schemas [14]. The integration method followed in the project is based on the work of [8, 10]. This framework was developed for the creation of web views, where each user has their own version of the same set of data. The framework developed by the authors has three ontology layers:

- **Local ontologies:** Relates local data sets to a desired domain. Each local ontology represents a database schema;
- **Exported views ontologies:** Intermediary level ontology to convert and map databases and concepts;
- **Domain ontology:** Shared ontology vocabulary that is related to the application and integrates all databases.

There are several initiatives that seek the semantic unification of different databases [15], [16]. Therefore, an alternative is to re-create the semantic context from database schemata and data documentation. Thus, the strategy is to build an ontological layer to facilitate access and reuse of information, both by humans and machines [17]. This is an approach adopted by both companies and population health entities, such as the WHO (World Health Organization) [15].

4 RESULTS AND EXPECTED RESULTS

It is expected that the integration of public health databases will help to map patient care pathways, predict public resource needs and minimize unnecessary spending. For this, the domain ontology of the proposed application was created and it has 200 axioms and is currently in testing. The concepts used to create the exported view ontology and the local ontology from SINASC were validated. The next steps are to create the exported view ontology and local ontology from SIM and SIH data sets and to validate them and their relations.

REFERENCES

[1] Shortliffe, E. H. & Cimino, J. J. *Biomedical informatics: computer applications in health care and biomedicine* (Springer Science & Business Media, 2013).

[2] Portal da saúde SUS: Transferência/download de arquivos. http://www2.datasus.gov.br/DATASUS/index.php?area=0901. Accessado em: 21-05-2018.

[3] Brasil, da Saúde, M., de Atenção a Saúde, S. & Departamento de Regulação, A. e. C. *Sistemas de Informação da Atenção à Saúde: Contextos Históricos, Avanços e Perspectivas no SUS* (Organização Pan-Americana da Saúde, 2015).

[4] Brasil, da Saúde, M. & da Estratégia e Saúde, C. G. *Estratégia de e-Saúde para o Brasil* (Ministério da Saúde, 2017).

[5] Silva, J. P. L. d., Travassos, C. M. d. R., Vasconcellos, M. M. d., Campos, L. M. *et al.* Revisão sistemática sobre encadeamento ou linkage de bases de dados secundários para uso em pesquisa em saúde no brasil. (2006).

[6] Souza, R. C. d. & Freire, S. M. Integração de dados ambulatoriais de quimioterapia e radioterapia nas bases de dados do SUS. 1904–1907 (2014).

[7] Maia, L. T. d. S., Souza, W. V. d., Mendes, A. d. C. G. & Silva, A. G. S. d. Use of linkage to improve the completeness of the SIM and SINASC in the Brazilian capitals. *Revista de saude publica* **51**, 112 (2017).

[8] Lopes, G., Vidal, V. & Oliveira, M. A framework for creation of linked data mashups: A case study on healthcare. In *Proceedings of the 22nd Brazilian Symposium on Multimedia and the Web*, 327–330 (ACM, 2016).

[9] Ekaputra, F. J., Sabou, M., Serral, E., Kiesling, E. & Biffl, S. Ontology-based data integration in multi-disciplinary engineering environments: A review. *Open J. Inf. Syst. (OJIS)* **4**, 1–26 (2017).

[10] Vidal, V. M. *et al.* Specification and incremental maintenance of linked data mashup views. In *International Conference on Advanced Information Systems Engineering*, 214–229 (Springer, 2015).

[11] Dong, X. L. & Naumann, F. Data fusion: resolving data conflicts for integration. *Proc. VLDB Endow.* **2**, 1654–1655 (2009).

[12] Wache, H. *et al.* Ontology-based integration of information-a survey of existing approaches. In *IJCAI-01 workshop: ontologies and information sharing*, vol. 2001, 108–117 (Citeseer, 2001).

[13] Psyllidis, A., Bozzon, A., Bocconi, S. & Bolivar, C. T. A platform for urban analytics and semantic data integration in city planning. In *International conference on computer-aided architectural design futures*, 21–36 (Springer, 2015).

[14] Doan, A. & Halevy, A. Y. Semantic integration research in the database community: A brief survey. *AI magazine* **26**, 83 (2005).