Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth
A. Ugon et al. (Eds.)

doi:10.3233/978-1-61499-852-5-6

Exploring Semantic Data Federation to Enable Malaria Surveillance Queries

Jon Haël BRENAS[a,1], Mohammad Sadnan AL MANIR[b], Kate ZINSZER[c], Christopher J.O. BAKER[b,d] and Arash SHABAN-NEJAD[a]
[a]The University of Tennessee Health Science Center (UTHSC) - Oak Ridge National Laboratory (ORNL) Center for Biomedical Informatics, Department of Pediatrics, Memphis, TN, USA, [b]University of New Brunswick, Saint John, New Brunswick, Canada, [c]Institut de recherche en santé publique de l'Université de Montréal, Montreal, Quebec, Canada, [d]IPSNP Computing Inc., Canada

Abstract. Malaria is an infectious disease affecting people across tropical countries. In order to devise efficient interventions, surveillance experts need to be able to answer increasingly complex queries integrating information coming from repositories distributed all over the globe. This, in turn, requires extraordinary coding abilities that cannot be expected from non-technical surveillance experts. In this paper, we present a deployment of Semantic Automated Discovery and Integration (SADI) Web services for the federation and querying of malaria data. More than 10 services were created to answer an example query requiring data coming from various sources. Our method assists surveillance experts in formulating their queries and gaining access to the answers they need.

Keywords. Interoperability, Web Services, Malaria Surveillance, Malaria Analytics, Distributed Data

Introduction

Malaria is an infectious disease caused by *Plasmodium* parasites and transmitted through the bites of an infected *Anopheles* mosquito. It is endemic throughout 91 tropical and subtropical countries and in 2015, it was estimated to have caused the death of 429,000 people [1] worldwide. Sub-Saharan African countries, however, bear the heaviest burden in the world and account for almost 90% of all registered malaria cases. The economic and societal costs of malaria are staggering, especially in the impoverished countries where it thrives [2]. The ability to control and eradicate malaria is part of the international goals for malaria community [3]. A key component in controlling and eradicating malaria is surveillance for the monitoring and evaluation of the trends, epidemiology, and interventions.

Malaria data is stored in distributed repositories, locally and globally, with various levels of granularity. For instance, Mapping Malaria Risk in Africa (MARA) [4] is an open-access Web-based platform designed to extract and display raw malario-metric data. Likewise, Zambia's District Health Information Software 2 (DHIS2) [5] is an open source software platform for reporting, analysis, and dissemination of data on

[1] Corresponding Author – E-mail: jhael@uthsc.edu

different health conditions. The goal of the Semantics, Interoperability, and Evolution for Malaria Surveillance (SIEMA) project [6] is to enable the federation and querying of distributed malaria data in a manner resilient to changes in data availability, structure and location. This paper introduces a central component of this solution, namely the use of semantic querying and data federation to access target information in distributed data repositories. SADI Semantic Web services [7] have been shown to improve semantic data access to a variety of programmatically accessible data repositories. In this paper, we present a deployment of SADI services and illustrate how surveillance experts can construct queries for malaria data without special regard to how the data is structured or distributed.

This paper is organized as follows: we start with a technical description of the SADI framework. This is followed by a brief overview of the registry of services and the graphical query client used in an example deployment. The paper concludes with some observations.

1. Methods

In this study, we have adopted a RESTful Web Service framework called the Semantic Automated Discovery and Integration (SADI) [7] which provides a set of conventions for creating Semantic Web services making publication of services easier. Services defined using SADI provide their I/O descriptions as well as a queryable semantic description of the service functionality. Service descriptions are archived in registries where they can be discovered by SADI query clients, SHARE [8] and HYDRA [9] using SPARQL as a query language. These clients can plan complex workflows and invoke services to retrieve target data in an automated fashion. The SADI framework uses RDF[S], OWL for data representation and modeling, and HTTP based recommendations (GET, POST) for interacting with the services. The SADI services receive and produce RDF instances of OWL classes. Input RDF data is annotated with service outputs to become an integrated instance of the output OWL class, thereby linking the input and output data.

SADI service input and output descriptions are authored using community adopted domain terminologies. For the current malaria surveillance study, we used terminologies from the Mosquito Insecticide Resistance Ontology (MIRO) [10]. In our trial deployment we used HYDRA, a commercial query engine for SADI services developed by IPSNP Computing Inc. In this paper, we illustrate HYDRA's intelligent graphical user interface that supports query composition by non-technical users. It has a keyword/natural language understanding capability that is employed to form a graph-based rendition of an end user's query that is then translated to SPARQL. HYDRA can interpret SPARQL and discover matches to SADI services stored in the registry. The SADI framework has already been used in other scenarios requiring complex data integration, such as clinical intelligence [9] and ecotoxicology [11], but never in the context of infectious disease surveillance. Use of the HYDRA's graphical interface for query composition as not been previously reported in the literature.

2. Malaria Services

Building services for the federation and querying of malaria data require consultation with end-users about the queries they want to be answered and the availability of relevant data sources. Figure 1 shows a registry of eleven SADI Semantic Web services, which are created to answer some queries in the field.

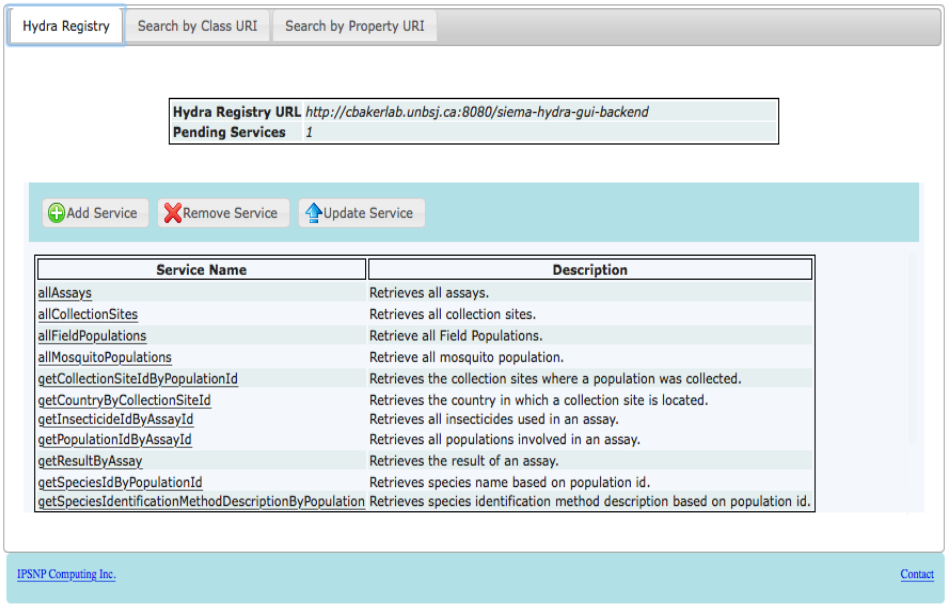

Figure 1. A screenshot of the registry of services used to answer the example query.

The declarative programming of input and output of the services define what a service expects as an input and what it produces as an output after the execution. Among the deployed services, there are four in the form of all*X*, which retrieve all information regarding *X* without expecting any input. They are: allAssays, allCollectionSites, allFieldPopulation, and allMosquitoPopulations. On the other hand, there are 7 services in the form of get*Y*By*Z*, which retrieve *Y* based on the input *Z*. They are:

getCollectionSiteIdByPopulationId, getCountryByCollectionSiteId, getInsecticideIdByAssayId, getPopulationIdByAssayId, getResultByAssay, getSpeciesNameByPopulationId, and getIdentificationMethodDescriptionByPopulationId.

For example, while the service allMosquitoPopulations retrieves all the identifiers of the mosquito populations, the service getSpeciesNameByPopulationId retrieves the name of a mosquito species for a given mosquito population identifier.

In this paper, we defined one query that we deemed interesting and created synthetic data that can be used to provide an answer. The query we used as an example was "What are the names of all the species of mosquito found in Uganda that are affected by the insecticide DDT?". The graph representation of this query can be found in Figure 2. This query returns the list of possible values of one variable, the species **?species** in the red variable-node. The species that correspond to the answer are the ones which are part of a population, identified by the blue concept-node MIRO_30000006, which was collected in a site in the country identified by the yellow

value-node containing the string *Uganda*. That population also had to be part of a bioassay during which a utilization of the insecticide named *Dichlorodiphenyltrichloroethane* was deemed to have a *positive* toxicological result. The services called to answer this query were allAssays, getCollectionSiteIdByPopulationId, getCountryByCollectionSiteId, getInsecticideIdByAssayId, getPopulationIdByAssayId, getResultByAssay and getSpeciesNameByPopulationId.

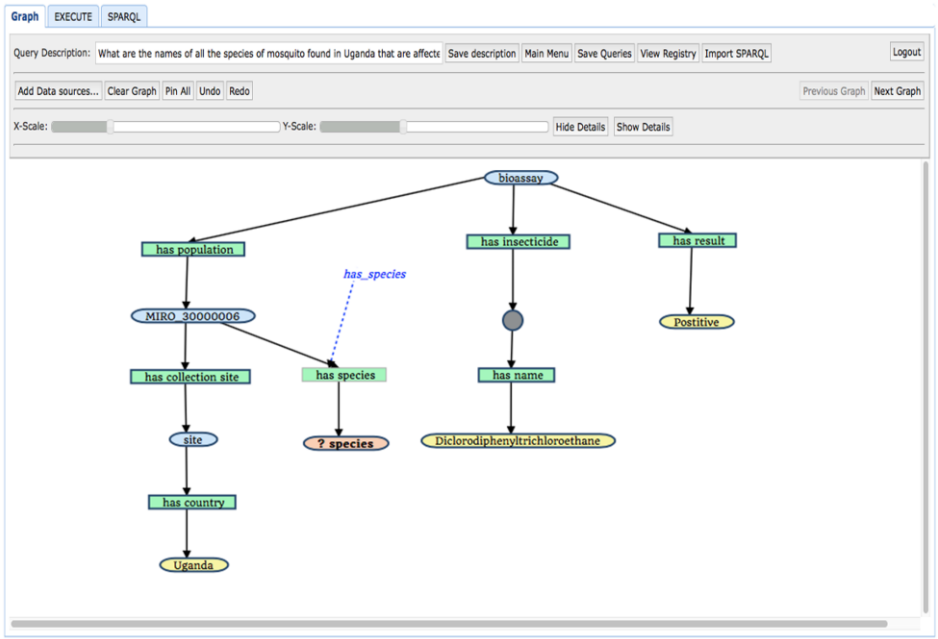

Figure 2. The graph representation of the query "What are the names of all the species of mosquito found in Uganda that are affected by the insecticide DDT?" The query specifically queries for the results of tests where a bioassay showed a positive toxic effect of the insecticide DDT on mosquitos collected from a site located in Uganda and select for the names of the mosquito species.

3. Conclusions and Discussion

In this work, we have introduced the use of a data integration platform that facilitates non-technical end-users to create complex queries over heterogeneously formatted and distributed data. The presented approach has been already shown beneficial in formulating and answering complex queries in other domains [9,11] including mission critical surveillance tasks. In [11] SADI has been used for querying information to aid detection of patients with hospital-acquired infections, such as Sepsis. Clinical guidelines stipulate diagnostic variables reflecting conditions such as fever, high white blood cell counts, deficiencies in oxygen reaching the body tissues and checking for these conditions is essential to the identification of identify patients at risk of infection. In recent work HYDRA GUI has made it possible for diagnostic variables to be readily translated into surveillance queries[2].

[2] http://tinyurl.com/IPSNP-HYDRA-Videos

Malaria surveillance practitioners are also interested in identifying incidents of infections, albeit at the population level over wide geographic regions. They also want to check the occurrence of environmental conditions that correspond with the spread of mosquito vectors, determine the impact of interventions and review the demographics of patients. In both scenarios, a framework for federation of heterogeneous distributed data and a tool for easy composition of multiple complex queries can greatly improve surveillance and facilitate interventions. In our current work**,** we are in the process of building and testing the appropriate resources (i) a registry of SADI services, (ii) a shareable library of saved SPARQL queries essential for malaria surveillance tasks. By recruiting SADI and HYDRA we seek to showcase a tangible approach for assembling surveillance routines composed of many complex queries across multiple data resources and formats. We anticipate that the example we provide will stimulate essential discussions in the community about strategies for surveillance tasks to the extent that a set agreed upon surveillance criteria can be realized and adopted.

Acknowledgments. Research supported by the Bill & Melinda Gates Foundation.

References

[1] WHO. World malaria report 2016. Technical report, 2016.
[2] J.L. Gallup and J. D. Sachs. The economic burden of malaria. American J. of Tropical Medicine & Hygiene, 2001.
[3] Roll Back Malaria Partnership. Rbm partnership strategic plan 2017-2020. http://rollbackmalaria.org/wp-content/uploads/2017/09/Draft-Strategic-Plan-as-at-14-Aug-2017.pdf. Accessed: 2018-02-07.
[4] Mapping malaria risk in africa. http://www.mara-database.org//login.html. Accessed: 2017-07-02.
[5] The dhis 2 web site. https://www.dhis2.org/. Accessed: 2017-06-27.
[6] J.H. Brenas, M.S. Al-Manir, C.J.O. Baker, and A. Shaban-Nejad. A malaria analytics framework to support evolution and interoperability of global health surveillance systems. IEEE Access, PP(99):16, 2017.
[7] M.D. Wilkinson, B. Vandervalk, L. McCarthy. The Semantic Automated Discovery and Integration (SADI) web service design-pattern, API and reference implementation. J. of Biomedical Semantics, 2(1):8, 2011.
[8] E. Dialynas, P. Topalis, J. Vontas, and C. Louis. Miro and irbase: It tools for the epidemiological monitoring of insecticide resistance in mosquito disease vectors. PLOS Neglected Tropical Diseases, 3(6):1–9, 06 2009.
[9] H. Boley. A RIF-Style Semantics for RuleML-Integrated Positional-Slotted, Object-Applicative Rules, pages 194–211. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.
[10] A. Riazanov, A. Klein, A. Shaban-Nejad, G. W. Rose, A. J. Forster, D. L. Buckeridge, and C. J. O. Baker. Semantic querying of relational data for clinical intelligence: a semantic web services-based approach. J. Biomedical Semantics, 4:9, 2013.
[11] A. Riazanov, M.M. Hindle, E.S. Goudreau, C.J. Martyniuk, and C.J.O. Baker. Ecotoxicology data federation with sadi semantic web services. In SWAT4LS, 2012.