Data Integration through Ontology-Based Data Access to Support Integrative Data Analysis: A Case Study of Cancer Survival

Hansi Zhang[1*], Yi Guo[1*], Qian Li[1], Thomas J. George[2], Elizabeth A. Shenkman[1], Jiang Bian[1§]

[1]Health Outcomes & Policy, University of Florida, Gainesville, Florida, USA
[2]The Division of Hematology and Oncology, Department of Medicine, University of Florida, Gainesville, Florida, USA

Abstract—**To improve cancer survival rates and prognosis, one of the first steps is to improve our understanding of contributory factors associated with cancer survival. Prior research has suggested that cancer survival is influenced by multiple factors from multiple levels. Most of existing analyses of cancer survival used data from a single source. Nevertheless, there are key challenges in integrating variables from different sources. Data integration is a daunting task because data from different sources can be heterogeneous in syntax, schema, and particularly semantics. Thus, we propose to adopt a semantic data integration approach that generates a universal conceptual representation of "information" including data and their relationships. This paper describes a case study of semantic data integration linking three data sets that cover both individual and contextual level factors for the purpose of assessing the association of the predictors of interest with cancer survival using cox proportional hazard models.**

Keywords—semantic data integration, ontology-based data access, cancer survival, integrative data analysis

I. Introduction

As the second leading cause of death, cancer is responsible for one in every four deaths in the US [1]. To improve cancer survival rates and prognosis, one of the first steps is to improve our understanding of contributory factors associated with cancer survival. Prior research has suggested that cancer survival is influenced by multiple factors from multiple levels. At the individual level, cancer survival is influenced by not only cancer stage of diagnosis and treatment, demographics, and financial status, but also risky health behaviors such as smoking, alcohol drinking, and physical inactivity. At the contextual level, cancer survival is influenced by public policies that influence health care delivery which could impact patients' travel distance to the treatment facility [2]. Prior epidemiologic research on cancer survival in the US has primarily focused on contributory factors from the individual level due to limited data availability. Most of these analyses used data from a single source, such as data from a hospital or a cancer registry. However, it is important to pool heterogeneous data sets for integrative data analysis (IDA) that simultaneously examine as many cancer survival predictors as possible (i.e. top down approach to model building) so that confounding effects among predictors can be fully understood.

Nevertheless, there are key challenges in integrating variables from different sources. Data integration is a daunting task because data from different sources can be heterogeneous in syntax (e.g., file formats, access protocols), schema (e.g., data structures), and semantics (e.g., meanings or interpretations). The effort required to connect different sources is substantial, especially due to the lack of clear definitions (i.e., data semantics) of variables and measures. Adopting a semantic data integration approach, we propose to generate a universal conceptual representation of "information" via ontologies to bridge syntactic, schematic, and semantic heterogeneities across different sources. The "information" includes data elements, modeled via common controlled vocabulary, and their semantic relationships. The use of ontologies can facilitate data integration in many ways, including metadata representation, automatic data verification, global conceptualization, and support for high-level semantic queries [3].

This paper describes a case study of semantic data integration linking three data sets that cover both individual and contextual level factors for the purpose of assessing the association of predictors of interest with cancer survival using cox proportional hazards models.

II. Methods

A. Data Sources

We obtained patients' demographic, tumor, treatment, and survival information from the 1996–2010 data of Florida Cancer Data System (FCDS), a statewide population-based registry supported by the Florida Department of Health and the Centers of Disease Control and Prevention. We obtained census tract-level poverty information from the 2000 U.S. census data, and also obtained 1996−2010 county-level smoking rates from the Behavioral Risk Factor Surveillance System (BRFSS) of the Centers for Disease Control and Prevention. All the underlying data sources are in a relational structure. Thus, we imported all of our source data into a relational database (i.e., MySQL).

B. The Process of Semantic Data Integration

Our approach for semantic data integration is based on an ontology-based data access (OBDA) [4] framework demonstrated in Fig 1. The first step of data integration is to construct synthesized, integrated descriptions (i.e., a global view) of the information coming from multiple heterogeneous sources, to provide users with a uniform query interface. Following the Global-as-View (GaV) approach for building data integration systems [3], a global ontology is built as a metadata representation of the data elements, their relationships within each source data schema as well as their relationships across different data sources. Then, semantic mappings need to be established between the global ontology and the data sources.

[*] Co-first authors: These authors contributed equally.
[§] Corresponding author: Jiang Bian, bianjaing@ufl.edu.

978-1-5090-3050-7/17/$31.00 ©2017 IEEE

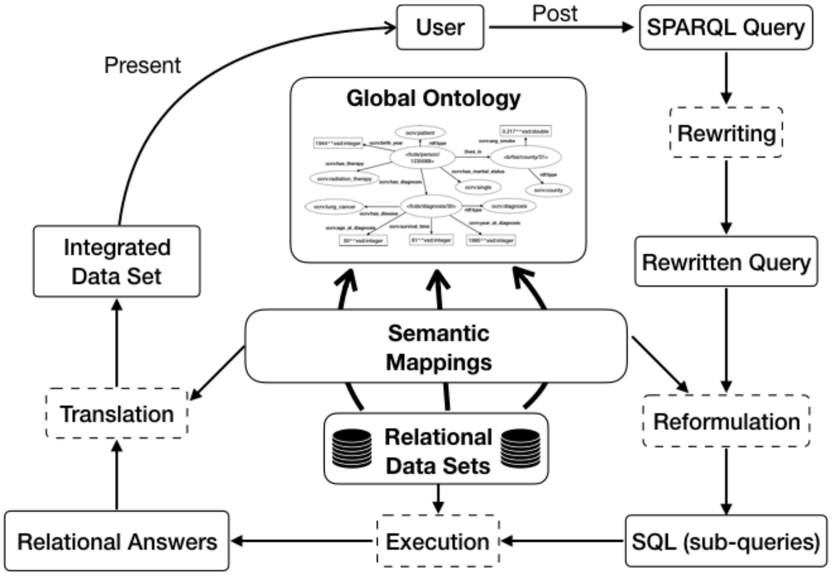

Fig. 1. The overall process of our semantic data integration approach through an ontology-based data access framework.

Given a conceptual view (i.e., the global ontology) of available data sources, a user can pose semantic queries for a data integration task (i.e., SPARQL, SPARQL Protocol, and RDF Query Language) over the global ontology. A semantic query is high-level as its formulation does not require one's awareness of source schemas. A high-level semantic query is re-formulated into a union of sub-queries over all the data sources, using the semantic mappings. The sub-queries are subject to the structure of source schemas, and often expressed in the native query languages of the sources (e.g., Structured Query Language, SQL commonly used for relational databases). The integration of the sub-query results constitutes the relational answer to the SPARQL query. After translation using the semantic mappings, the integrated data set is presented to the user. The global ontology and the semantic mappings are constructed manually according to our integrative data analysis use cases.

C. Implementation of Semantic Data Integration with Ontop

To develope an ontology-driven semantic data integration approach, we used the Ontop platform that allows for semantic queries against relational databases [5]. Within the Ontop system, data elements (and the relationships between the data elements) in the source data were mapped to an ontology and presented as virtual Resources Description Framework (RDF) graphs. Subsequently, the virtual RDF graphs can be queried with SPARQL. We used Protégé to construct a global ontology, namely the Ontology for Cancer Research Variables (OCRV), using the Web Ontology Language (OWL) and the Ontop Protégé plugin.

1) Construct OCRV and Design the Semantic Mappings. Construct classes, properties, and relations: In general, a class is an abstraction for a group of individuals. Object properties represent the relationships between individuals, and the datatype properties link individuals to data values (i.e., to describe the attributes of an individual). In our work, classes are concepts such as "patient", "disease", and "diagnosis". When defining these classes, we used the ontological resources in the BioPortal [6] as a foundation for modeling, re-using the preferred terms and definitions that are used in existing widely accepted ontologies.

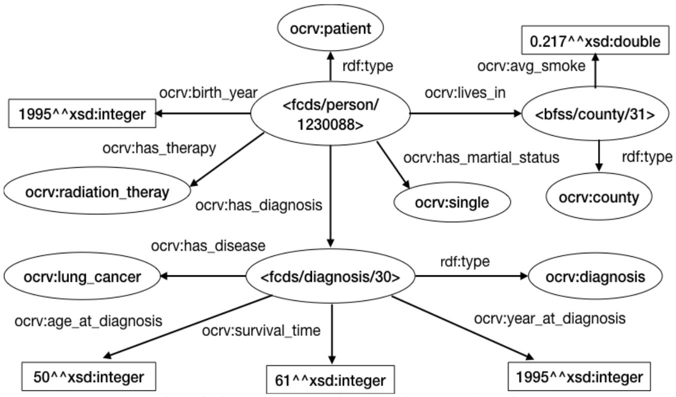

Fig. 2. An example of data annotated with the ontolog for cancer research variables (OCRV).

Design mappings: In Ontop, an OBDA model is needed to describe how the data in the data sources are related to the global ontology. In particular, mapping axioms are used in an OBDA model to map data in the source databases into a set of ABox assertions/RDF triples. Each mapping axiom consists of three parts: MAPPING_ID, SOURCE and TARGET. The MAPPING_ID is a unique identity for each mapping axiom, the SOURCE is an SQL query to retrieve relevant data from the database and the TARGET is a RDF triple template. These mapping axioms can be divided into three basic scenarios, as described below and demonstrated in Fig. 2.

a) Class instance: This kind of mapping axioms is used to declare that a set of individuals are of a certain type through ***rdf:type***. For exmaple, in TABLE 1, the SQL query in the SOURCE extracts all patients' identifiers (i.e., PATIENT_ID) from the FCDS database table. The triples in the TARGET asserts that the individuals indentified by the these identifiers (i.e., ***:fcds/person/{PATIENT_ID}***) represent instances of the class ***ocrv:patient***.

TABLE 1. AN EXAMPLE OF MAPPING CLASS INSTANCES.

SOURCE	SELECT `PATIENT_ID`
	FROM `fcds_sample`

TARGET	:fcds /person/{PATIENT_ID} a ocrv:patient .

a is an abbreviation for ***rdf:type***.

b) Datatype properties: The general idea of mapping datatype properties in Ontop is to link attributes of individuals to columns in the database tables. We created 13 datatype properties in OCRV based on our data analysis use cases such as ***ocrv:date_of_diagnosis***. TABLE 2 shows an example of representing the birth years (i.e., data values) of our FCDS patients through the datatype property ***ocrv:birth_year***.

TABLE 2. AN EXAMPLE OF MAPPING DATATYPE PROPERTIES.

SOURCE	SELECT `PATIENT_ID`, `Birth_Year_N240`
	FROM `fcds_sample

TARGET	:fcds/{PATIENT_ID} a ocrv:patient ; ocrv:birth_year {Birth_Year_N240}^^xsd:integer .

c) Object properties: Object properties in Ontop are used to link between individuals and/or classes. We created 9 object properties such as ***ocrv:has_stage***, ***ocrv:has_race***, ***ocrv:has_ethnicity***, and ***ocrv:has_marital_status***. TABLE 3 shows an example of linking patients with their counties of residency through the object property ***ocrv:lives_in***.

TABLE 3. AN EXAMPLE OF MAPPING OBJECT PROPERTIES.

SOURCE	SELECT `PATIENT_ID`,
	`County_at_DX_N90`
	FROM `fcds_sample

TARGET	:fcds /{PATIENT_ID} a ocrv:patient ; ocrv:lives_in :brfss /county/{County_at_DX_N90}

2) Build Semantic Queries. After establishing the mappings between the global ontology (i,e., OCRV) and the underlying relational databases, Ontop is able to realize the relational data into virtual RDF graphs on-the-fly. We can then use semantic queries, expressed in SPARQL, to retrieve and manipulate the data stored in these RDF graphs. Based on the data integration needs, our queries can be classified into four categories: (1) queries that extract variables directly linked to a patient without the need for any processing; (2) queries that are used to link a patient to environmental factors through geographic variables; (3) queries that need to leverage the relationship between classes to generate the desired results; (4) queries that need to prepocess the raw data to produce the desired results. We will discuss these query use cases in detail in the RESULT section.

3) Create a Data Integration Pipeline. The goal of our data integration tasks was to link different data sources to form a single pooled data set for statistical models. Thus, our final step was to assemble SPARQL queries to produce and format the needed data in an automated fashion. As requried by data analysis models, the format of the pooled data set was organized into a data table (i.e., a matrix), where each row represents a patient, and each column represents a risk factor. We wrote a data integration pipeline in Java using the Ontop OWL API. We generated the data table column by column based on the 4 types of SPARQL queries we described above.

III. RESULT

We present results for the 4 types of SPARQL queries and an example result of the data integration pipline.

A. The Four Types of Semantic Queries

1) The first type of queries was those used to extract variables directly linked to a patient without further data processing need. For example, in the IDA, we needed a column to represent whether a patient had raditaion therapy or not. To do so, our process was to use a SPARQL query to find all patients who had raditaion therapy and marked them with "1" in the "has_radiation_therapy" column of the final data table. Fig. 3. shows how a patient's radiation therapy information is modeled. TABLE 4 shows the corresponding SPARQL query, where ***?p*** represents patients. Note In SPARQL, variable names are prefixed with the question mark ("?") symbol.

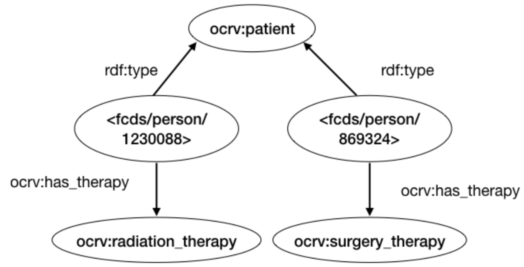

Fig. 3. The relationships between patients and therapies in OCRV.

TABLE 4. A SPARQL QUERY FOR LISTING PATIENTS WHO HAD RADIATION THERAPIES.

```
PREFIX :<http://www.semanticweb.org/ontologies/OCRV#>
SELECT ?p where {
   ?p a ocrv:patient.
   ?p ocrv:has_therapy ocrv:radiation_therapy. }
```

2) The contextual and environmental factors used in cancer survival models were linked to indiviudal patients through their residency. In our FCDS data, each patient's residency was recorded using both a census tract code and county code. The average smoking rate of each county (i.e., represented by a county code) was obtained from the BRFSS. Our task was to extract the average county smoking rate and link it to each individual patient through common county codes. The SPARQL query for this task is shown in TABLE 5. The variables ***?p*** and ***?y*** represent instances of ***ocrv:patient*** and ***ocrv:county***, respectively; which were linked by the object property ***ocrv:lives_in***. Similarly, we used ***ocrv:avg_smoke***, a datatype property, to link individuals of ***ocrv:county*** with the corresponding counties smoking rates. These relationshps were demonstrated in the OCRV as shown in Fig. 2.

TABLE 5. QUERY THE COUNTY LEVEL SMOKING RATES BASED ON RESIDENCY.

SPARQL query:

```
PREFIX :<http://www.semanticweb.org/ontologies/OCRV#>
SELECT ?p ?y ?z WHERE {
   ?p a ocrv:patient.  ?p ocrv:lives_in ?y.
   ?y a ocrv:county.  ?y  ocrv:avg_smoke ?z}
```

Result:

?p (ocrv:patient)	**?y (ocrv:county)**	**?z (ocrv: avg_smoke)**
<fcds/person/869324>	<brfss /county/31>	0.217

3) In our data analysis use cases, the raw categorical variables were grouped into different subgroups, a common practice in building prediction models. To produce desired grouping, we encoded the grouping knowledge into OCRV ontology, leveraging object properties to represent these grouping logics. For example, there were 7 different categories for marital status (i.e., "single", "divorced", "widowed", "separated", "married", "unknown" and "unmarried") in the FCDS data. In the cox model, "single", "divorced", "widowed", "separated", and "unmarried" were considered "single". Thus, in OCRV, we modeled ***ocrv:divorced***, ***ocrv:widowed***, ***ocrv:separated***, and ***ocrv:unmarried*** as subclasses of ***ocrv:single***. An example of how we moleded marital status in OCRV is shown in Fig. 4.

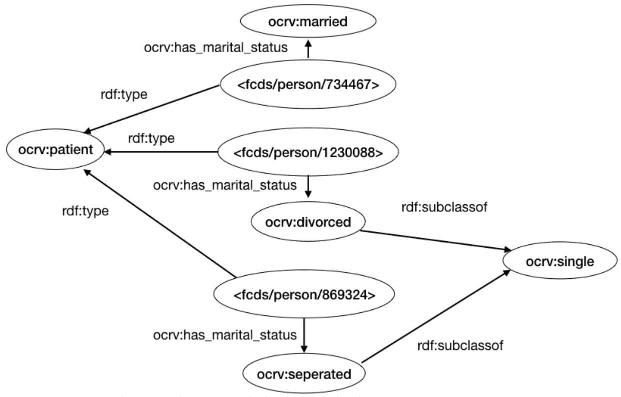

Fig. 4. Representing patients and their marital statuses.

With this ontological structure, the SPARQL query to retrieve patients' marital status according to the required categorization was rather straightforward as shown in TABLE 6.

TABLE 6. A SPARQL QUERY FOR LISTING PATIENTS WHOSE MARTIAL STATUS IS SINGLE.

```
PREFIX :<http://www.semanticweb.org/ontologies/OCRV#>
SELECT DISTINCT ?p WHERE {
   ?p a ocrv:patient; ocrv:has_marital_status ocrv:single }
```

4) Some variables needed in the data analysis are derived from the raw source data. For example, in our cancer survival models, we only needed the year of cancer diagnosis, whereas the raw FCDS data recorded the date of diagnosis in the format of "mm/dd/yyyy". Thus, for this type of queries, the raw data needed to be manipulated (e.g., through transformation, calcuation, or extraction) to generate the desired variables for data analysis. TABLE 7 illustrates how year of diagnosis was extracted from date of diagnosis through a SPARQL query.

TABLE 7. A QUERY FOR RETRIEVING PATIENTS' YEAR OF DIAGNOSIS.

SPARQL query:

```
PREFIX :<http://www.semanticweb.org/ontologies/OCRV#>
SELECT ?p ?year WHERE {
   ?p a ocrv:dianoses.
   ?p ocrv:date_of_diagnosis ?d BIND(str(year(?d)) AS ?year)}
```

Result:

?p (ocrv:patient)	?year(ocrv:date_of_diagnosis)
<fcds/person/869324>	1994

B. The Semantic Data Integration Pipeline

With all the necessary SPARQL queries clearly specified, creating the data integration pipeline in Java is straightforward with the Ontop OWL API. A snippet of the final integrated data set generated through combining the 4 different types of SPARQL queries is shown in TABLE 8.

TABLE 8. A SAMPLE RESULT GENERATED FROM THE SEMANTIC INTEGRATION PIPLINE.

ID	year of diagnosis	marital status	average smoke	radiation therapy
1230088	1944	single	0.217	1
…	…	…	…	…
869324	1930	single	0.185	0

IV. CONCLUSIONS

In this study, we demonstrated the feasibility and advantages of using an ontology-based semantic data integration approach to link heterogeneous data sources to create a pooled data set of IDA. With a semantic data integration approach, many data processing needs and knowledge can be encoded in the ontology, and thus data analysts no longer need to worry about the syntactic, schematic, and semantic heterogeneities in data from different sources.

ACKNOWLEDGMENT

This work was supported in part by NIH grants UL1TR001427, the OneFlorida Cancer Control Alliance (funded by James and Esther King Biomedical Research Program, Florida Department of Health Grant Number 4KB16), the OneFlorida Clinical Research Consortium (CDRN-1501-26692) funded by the Patient Centered Outcomes Research Institute (PCORI). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or PCORI.

REFERENCES

[1] "CDC - Statistics for Different Kinds of Cancer," 27-Jun-2017. [Online]. Available: https://www.cdc.gov/cancer/dcpc/data/types.htm. [Accessed: 26-Sep-2017].

[2] M. W. Vetterlein *et al.*, "Impact of travel distance to the treatment facility on overall mortality in US patients with prostate cancer," *Cancer*, vol. 123, no. 17, pp. 3241–3252, Sep. 2017.

[3] H. Xiao, "Query Processing for Heterogeneous Data Integration Using Ontologies," University of Illinois at Chicago, Chicago, IL, USA, 2006.

[4] H. Wache *et al.*, "Ontology-based Integration of Information - A Survey of Existing Approaches," in *Proceedings of IJCAI-01 Workshop: Ontologies and Information Sharing*, Seattle, WA, pp. 108–117.

[5] D. Calvanese *et al.*, "Ontop: Answering SPARQL queries over relational databases," *Semantic Web*, vol. 8, no. 3, pp. 471–487, 2017.

[6] National Center for Biomedical Ontology, "BioPortal," *National Center for Biomedical Ontology*, 2005-2017. [Online]. Available: http://bioportal.bioontology.org/. [Accessed: 11-Feb-2017].