Journal of Biomedical Informatics 42 (2009) 1035–1045

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Integration of prostate cancer clinical data using an ontology

Hua Min *, Frank J. Manion, Elizabeth Goralczyk, Yu-Ning Wong, Eric Ross, J. Robert Beck

Fox Chase Cancer Center, 333 Cottman Avenue, Philadelphia, PA 19111, USA

ARTICLE INFO

Article history:
Received 19 January 2009
Available online 2 June 2009

Keywords:
Data integration
Ontology
Knowledge sharing
Semantic heterogeneity

ABSTRACT

It is increasingly important for investigators to efficiently and effectively access, interpret, and analyze the data from diverse biological, literature, and annotation sources in a unified way. The heterogeneity of biomedical data and the lack of metadata are the primary sources of the difficulty for integration, presenting major challenges to effective search and retrieval of the information. As a proof of concept, the Prostate Cancer Ontology (PCO) is created for the development of the Prostate Cancer Information System (PCIS). PCIS is applied to demonstrate how the ontology is utilized to solve the semantic heterogeneity problem from the integration of two prostate cancer related database systems at the Fox Chase Cancer Center. As the results of the integration process, the semantic query language SPARQL is applied to perform the integrated queries across the two database systems based on PCO.

1. Introduction

Biomedical data systems and their associated information models, terminologies, protocols, and data dictionaries have often been developed independently. Consequently, data integration has become an important tool for biomedical researchers since the data obtained from experiments frequently needs to be combined with the data or the annotations derived from other systems. For example, it will significantly increase the value when human tissue specimens are combined with the medical data describing morphology, histopathology, and so on. Further, the integration of the data from a variety of sources improves the clinical decision making process and the quality of patient care [1,2]. Data integration has the capacity to expand research scope by focusing on common elements across studies and thus, create opportunities to achieve large enough sample sizes to detect the significant results.

Current data integration research is concerned with the semantic integration problem. The problem involves the best approach to resolve semantic conflicts (the conflict caused by using different terms in the heterogeneous systems to express the same entity in reality) among the heterogeneous data sources. A common strategy to address the semantic conflicts is through the use of an ontology with explicitly defined schema terms [3,4]. This approach is called ontology-based data integration.

In this paper, we demonstrate the implementation of the Prostate Cancer Information System (PCIS) based on one of the ontology-based data integration approaches. Besides providing services for researchers and clinicians in the field, it also provides a verified solution for large biomedical data integration projects (e.g., cancer Biomedical Informatics Grid (caBIG) [5,6], etc.). PCIS demonstrates the advantages of the ontology-based data integration strategy presented in Section 1.4.2.

1.1. Ontology

The term ontology has its origin in philosophy, where it is the name of a fundamental branch of metaphysics concerned with being or existence. In both computer and information science, ontology is a data model that represents a set of concepts within a domain and the relationships among these concepts. In short, ontology is a specification of a conceptualization [7]. A conceptualization is an abstract, simplified view of the world. The specification is a declarative representation of the conceptualization in a concrete form. It tries to interpret knowledge in a way that a computer can process unambiguously and consequently encode the concepts and relationships in a computer-usable language. The Web Ontology Language (OWL) is recommended by W3C to represent the web ontologies [8]. OWL has a greater machine interoperability for the web content than the Extensible Markup Language (XML), DARPA Agent Markup Language (DAML), Resource Description Framework (RDF) [9], and RDF Schema. OWL has three sublanguages: OWL Lite, OWL DL, and OWL Full [8]. OWL DL is chosen as the ontology representation language in this paper.

A number of ontologies have already been developed in the disciplines of biomedicine. The Unified Medical Language System (UMLS) is a comprehensive source of biomedical terminology that

* Corresponding author. Address: Population Science Department, Fox Chase Cancer Center, 333 Cottman Avenue, Philadelphia, PA 19111, USA. Fax: +1 215 728 2553.
E-mail address: hua.min@fccc.edu (H. Min).

1532-0464/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2009.05.007

consists of a large number (over 100) of national and international vocabularies and classifications [10]. The National Center for Biomedical Ontology (NCBO) [11] builds a library of biomedical ontologies known as the Open Biomedical Ontologies (OBO) [12] which is now comprised of more than 70 biomedical ontologies (such as Cell Ontology (CO) [13], MGED Ontology [14], Foundational Model of Anatomy (FMA) [15], and National Cancer Institute Thesaurus (NCI Thesaurus) [16,17], etc.). The Ontology for Biomedical Investigations (OBI) [18] project has developed an integrated ontology for the description of life-science and clinical investigations. The OBI Consortium is a member of the OBO Foundry. Currently, the OBI uses the Basic Formal Ontology (BFO) [19] as its upper-level ontology. The upper-level ontology captures mostly concepts that are basic to human understanding of the world. It describes very general concepts that are the same across all domains.

1.2. Ontology application in the database and information systems

In the field of database and information systems, ontology plays a crucial role in integrating the data from multiple heterogeneous resources by transforming the underlying data into a common representation and transmitting this knowledge to the application programs [20,21]. The semantic heterogeneity has been identified as the most challenging issue of data integration since it requires understanding of the relationships between the data and the real world objects, often based on various points of the view. Ontology provides a solution to address the semantic heterogeneity problem [4]. It provides formal definitions of the terms used in the data sources, and renders the implicit meaning of the relationships among the different terminologies of the data sources explicitly. For example, one can determine whether two classes of the data items from two different database systems are equivalent, or whether one is a subset of another by using ontology. Ontology also allows the users to query different database systems as one by tying them together at a semantic level.

1.3. Prostate cancer data at Fox Chase Cancer Center (FCCC)

Prostate cancer is the most common cancer (excluding skin cancer) among American men. American Cancer Society estimated that there would be more than 186,320 new cases of prostate cancer and approximately 28,660 deaths in 2008 in the United States [22].

Fox Chase Cancer Center (FCCC) is a National Cancer Institute designated comprehensive cancer center which admits over 7000 new patients annually. Given its high incidence of prostate cancer, Fox Chase has a very large number of prostate cancer patients who are currently under treatment or have completed treatment. There are several major information systems at FCCC that contain the data relevant to prostate cancer. These include the Tumor Registry, Pathology Report System, Pharmacy, Risk Assessment Program, and Laboratory System, etc. These systems were designed for different tasks, and were either purchased or developed independently. Consequently, these systems use different underlying data and information models, as well as lexicons and vocabularies to represent the data captured at the point of service. For example, the systems use diverse staging standards to describe how far the cancer has spread. The Tumor Registry at FCCC uses the American Joint Committee on Cancer (AJCC 6 Edition) TNM staging system [23]. However, the Radiation Oncology Department at FCCC employs the FIGO staging system [24]. A translation between the two staging systems must be applied to query the data across these two database systems. The development of a common interface for information retrieval will benefit the efforts in prostate cancer prevention, diagnosis, treatment, and research for improving the quality of prostate cancer patient care.

1.4. Data integration methods

Biomedical data are collected from a large range of various fields including daily clinical practice, clinical trials, and scientific experiments. Data integration is the process of combining the data residing in different data sources to provide the users with a unified query interface to access these data [25]. The traditional data integration methods include data warehouse and database federation. Besides these two traditional methods, the ontology-based data integration has attracted attention because of its ability to address the semantic heterogeneity problem.

1.4.1. Data warehouse and database federation

A data warehouse is a repository of an organization's electronically stored data [26]. Data are extracted, transformed, and loaded (ETL) into a central repository from different sources. Besides the data itself, the data warehouse contains methods to retrieve and analyze data, to extract, transform and load data, and to manage it. The advantages of the data warehouse include improving users' ability to access a wide variety of data sources and increasing the data consistency. The data warehouse has a very high reliability and faster query response time since all data are located in the central data repository. It also has several disadvantages. The initial cost for the data warehouse is high since all data sources need to be transformed and copied into the central repository. The data warehouse needs to be refreshed periodically (e.g., daily or weekly) since it can be outdated relatively quickly.

A federated database system is considered as a meta-database management system that transparently integrates multiple autonomous database systems into a single conceptual view of the integrated database. The data sources are considered "federated" because the data is not copied into a central repository, rather, the federation server maintains indices or links to the relevant records or data of interest in the source systems [27]. The data sources are interconnected via a computer network, and may be geographically decentralized. In essence then, the database federation works as a virtual data warehouse [28,29]. In addition to the benefits of the data warehouse, the database federation provides accessibility to the live data and functions. The cost for constructing and maintaining a database federation is lower than that of a data warehouse since there is no need to gather information into a central repository. Security is sometimes considered enhanced, as the original data owners maintain control over authorization to the data contained in their systems. However, the query performance for the database federation is limited by a number of factors: network configuration and performance, schema design, and the availability of the source database systems.

1.4.2. Ontology-based data integration

The ontology-based data integration involves the use of ontology to effectively combine the data and/or information from the multiple heterogeneous sources. It provides a semantic layer on the top of the underlying data. The primary goal of the ontology is to provide a set of mechanisms for solving the semantic heterogeneity problems.

Ontology-based data integration has unique advantages [3]. It has a stable conceptual interface to the database systems because the ontology provides a rich and predefined vocabulary. The conceptual interface is independent of the database schemas. The knowledge represented by the ontology can be utilized to translate the relevant data sources into a common frame of the reference. The ontology-based data integration supports consistent management and recognition of the inconsistent data. It also provides a mechanism to define queries based on the concepts of the ontology and present the query results in a unified and structured form. The ontology-based data integration also brings the challenges for do-

main experts and computer scientists. Domain experts need to construct, merge, and maintain the domain ontologies. Each dataset needs to be registered (linked) to the ontology. Computer scientists have to build an integrated system based on the ontological registered datasets.

In ontology-based data integration, it requires a mediator system to represent all objects in the domain of the interest. Queries are then directed against the mediator, which in turn deals with the details of querying the source systems. This model of data integration is classified as a Local-as-View [30], to denote that queries on the local (source) databases are reformulated in terms of the global mediation. While other models are possible, Local-as-View models have been shown to scale better and be easier to maintain.

Variant approaches have been developed for the ontology-based data integration. They can be classified into ontology-based data warehouse or ontology-based database federation. Though, ontology is utilized in both approaches, each approach is distinguished by the data location and ontology construction. The CoryneRegNet [31] is one of the examples for the ontology-based data warehouse. The CoryneRegNet is designed to facilitate the genome-wide reconstruction of transcriptional regulatory networks of corynebacteria and *Escherichia coli*. The data related to transcriptional regulation from different sources are first imported into a single data repository. Then, the data model of the data repository is converted to ontology-based data structure. The ontology-based database federation has been applied to integrate two neuroscience databases (NeuronDB and CoCoDat) [32]. The D2RQ [33] is applied to translate the schema of each database into the corresponding OWL ontology. The two ontologies are merged by the "SameAs" construct in OWL which relates the equivalent concepts of the two ontologies.

As a proof of concept, we present a web-based Prostate Cancer Information System (PCIS). PCIS is applied to demonstrate the ontology-based data integration approach for the integration of two prostate cancer related databases at FCCC. The ontology is constructed based on the domain knowledge rather than the database schemas in PCIS. It demonstrates the necessary to introduce the properties of the concepts to meet the requirements from the ontology-based data integration. The data stored in the database systems are mapped into the corresponding concept and its properties of the global ontology. The drawback of this approach is that the ontology construction does not utilize the knowledge existing inside database schemas. It also needs external resources (such as domain experts, etc.) and greater effort to construct the ontology.

2. Methods

The detailed ontology-based data integration methods are presented in this section. These methods are utilized to integrate a series of the observations for the prostate cancer data from the FCCC radiation therapy outcome database, as well as the Tumor Registry. The system architecture of PCIS is presented in subsection 2.1. The data contents are described in subsection 2.2. The Prostate Cancer Ontology, the mappings between the ontology and the database systems, and the semantic web query are presented in the rest of the subsections sequentially.

2.1. System architecture of PCIS

The architecture of PCIS is presented in Fig. 1. Based on the functionalities, the system is divided into three subsystems: (1) Data storage, (2) Mapping, and (3) Data query.

The data storage subsystem consists of several independent database systems. As a proof of concept, PCIS only contains two database systems, Prostate Cancer Database and Tumor Registry. They are designed independently from each other, and are used for different tasks by independent groups of the users. The data are collected, stored, and maintained in each database system separately. Each database system contains the information of the prostate cancer patients. Since they serve the users from the different departments, the contents of each database system may be updated simultaneously and independently.

The mapping subsystem contains the mapping server to perform the functionalities of the mediator system. The mapping server stores the mappings and the genetic conversation functions between the ontology and the database systems. It also publishes the contents of the database systems on the Semantic Web. The declarative language D2RQ [33] is applied to describe the mappings between the relational database schemas and the OWL/RDFS ontologies. It generates the mapping file from table structures of the database systems in PCIS. Then, the mapping file is customized by replacing the auto-generated terms with the terms of the Prostate Cancer Ontology (PCO). The D2R server [34] serves as the mapping server in PCIS. It applies the mapping file to publish the contents of the relational database systems on the Semantic Web. The D2R server provides the functionalities to browse and search the RDF. Since RDF is the representation of the database systems, the D2R server can be applied to browse and search these systems.

The data query subsystem consists of the SPARQL (a query language for RDF [35]) interface which allows users to submit the queries. The interface enables users to search and query the database systems using the SPARQL query language over the SPARQL protocol. The web application provides a user friendly interface. The detailed description of each component in PCIS will be presented in the following subsections.

2.2. Data sources

There are several database systems that contain the prostate cancer information at FCCC. Two of them are selected as the data sources for PCIS. The first one is the Prostate Cancer Database which is operated and maintained by the radiation oncology department at FCCC. Currently, it has approximately 5000 records for prostate cancer patients who are treated at the radiation oncology department. The second one is FCCC's Tumor Registry system that maintains demography, cancer specific treatments, and follow-up for the patients with a reportable neoplasm. There are about 3000 records for the prostate cancer patients in the Tumor Registry. The Prostate Cancer Database focuses on the detailed treatment information while the Tumor Registry focuses on the summary of the patients. The contents overlap in some degree between the two database systems such as demography, medical history, and diagnosis.

2.3. Prostate Cancer Ontology (PCO)

OWL-DL is utilized to construct PCO for PCIS using the ontology editor tool, Protégé 3.3 [36]. PCO provides the common, shared, and formal description of the important concepts, relationships, and properties/attributes for prostate cancer. PCO is developed by merging the concepts from two commonly used vocabularies and ontologies, NCI Thesaurus and FMA. PCO inherits the concepts of prostate cancer from NCI Thesaurus. It also inherits the concepts of prostate anatomical structure from FMA. The new concepts identified from the database systems are added into PCO if they are not covered in either NCI Thesaurus or FMA. For example, the CNTO-328 is a monoclonal antibody to IL-6 used in one of the prostate cancer clinical trials at FCCC. It is added as one of the children for the concept "Drug in Clinical Trial" in PCO. The properties for

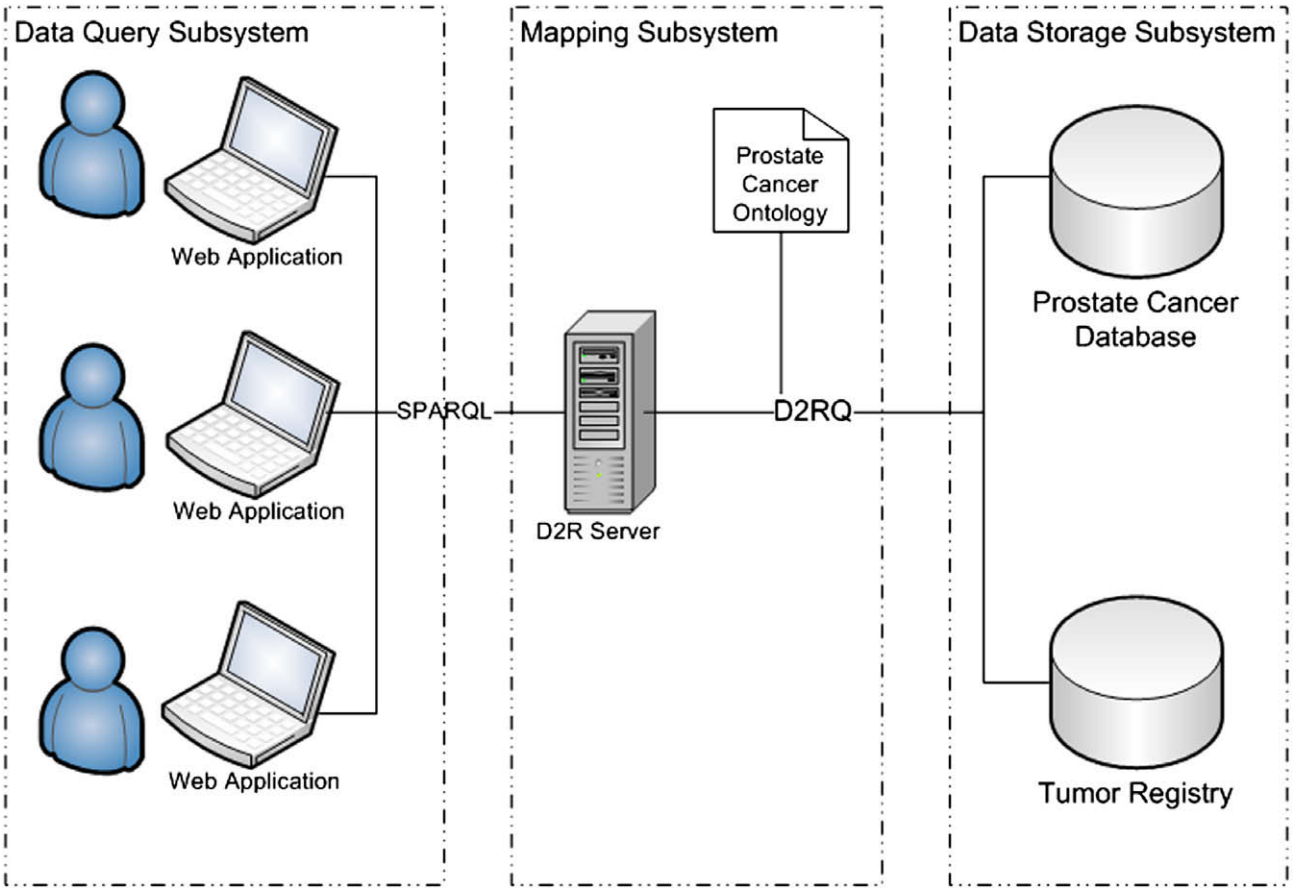

Fig. 1. Prostate Cancer Information System (PCIS) architecture.

each concept are introduced to PCO too. These properties are determined by the database systems. They are applied to annotate the data from the database systems.

2.4. Mappings between database schemas and ontology

In the mapping subsystem, D2R server follows the mapping file to browse and search the data. The mapping file contains the mappings between the database schemas and the ontology. The mapping process links each component (table, column, and constraint) of the Prostate Cancer Database and the Tumor Registry to its corresponding component (concept, property, or relationship) of PCO.

A portion of the mappings between PCO and two database schemas is presented in Fig. 2. The hierarchical structure of PCO is presented in the rectangle A on the top of the figure. The hierarchical (is-a) relationship among the concepts of PCO are shown as the arrows in the figure. The is-a relationship connects a more specific concept (child concept) to a more general concept (a parent). It serves as the ontology's backbone and supports the property inheritance. The small solid rectangle inside the rectangle A represents the concept of PCO. The properties for each concept are presented as ovals connected to each concept in the figure too. The properties are inherited from the parent to its children along the is-a hierarchy. New properties are also introduced by the concepts. For example, patient inherits properties such as name, sex and age from its parent Person. It also introduces a new property, MRN (Medical Record Number), which is patient specific (see Fig. 2).

The sample tables of the two database systems are presented in rectangle B and rectangle C. In rectangle B, V_DEMOP_PRCA_RADONC and BX are two tables of the Prostate Cancer Database. These tables contain patient's demography and biopsy information, respectively. In rectangle C, ORA_PT and ORA_DG are two tables of the Tumor Registry. They describe the information about patient's demography and diagnosis, respectively. The mappings are built between the tables of these two database systems and the concepts of PCO (shown as bold dotted double-headed arrows in the figure). For example, the table BX in the Prostate Cancer Database is mapped to its corresponding concept Biopsy of PCO. The tables containing the same information are mapped to the same concept even if they may have different names. For example, V_DEMOP_PRCA_RADONC in the Prostate Cancer Database and the ORA_PT in the Tumor Registry are mapped to the same concept Patient. The data in the relational database systems are mapped to the semantic layer through these mappings. These data are linked together thus, allowing PCIS to provide the integrative services.

Several software tools (such as DataMaster [37], R2O [38], RDB2Onto [39], and D2RQ) are available to perform the mapping tasks. They can automatically generate the mapping file from the table structures of the database systems to their corresponding ontological structures. The D2RQ tool is selected to create the mapping file for PCIS. The mapping file is customized by replacing the auto-generated terms with those of PCO. A sample of the customized mapping file is presented in Fig. 3. It shows the mappings between the V_DEMOP_PRCA_RADONC table and its corresponding concept Patient of PCO. The mappings between database tables and their corresponding concepts are created by using the command d2rq:ClassMap. The command d2rq:PropertyBridge maps the table columns to their corresponding prop-

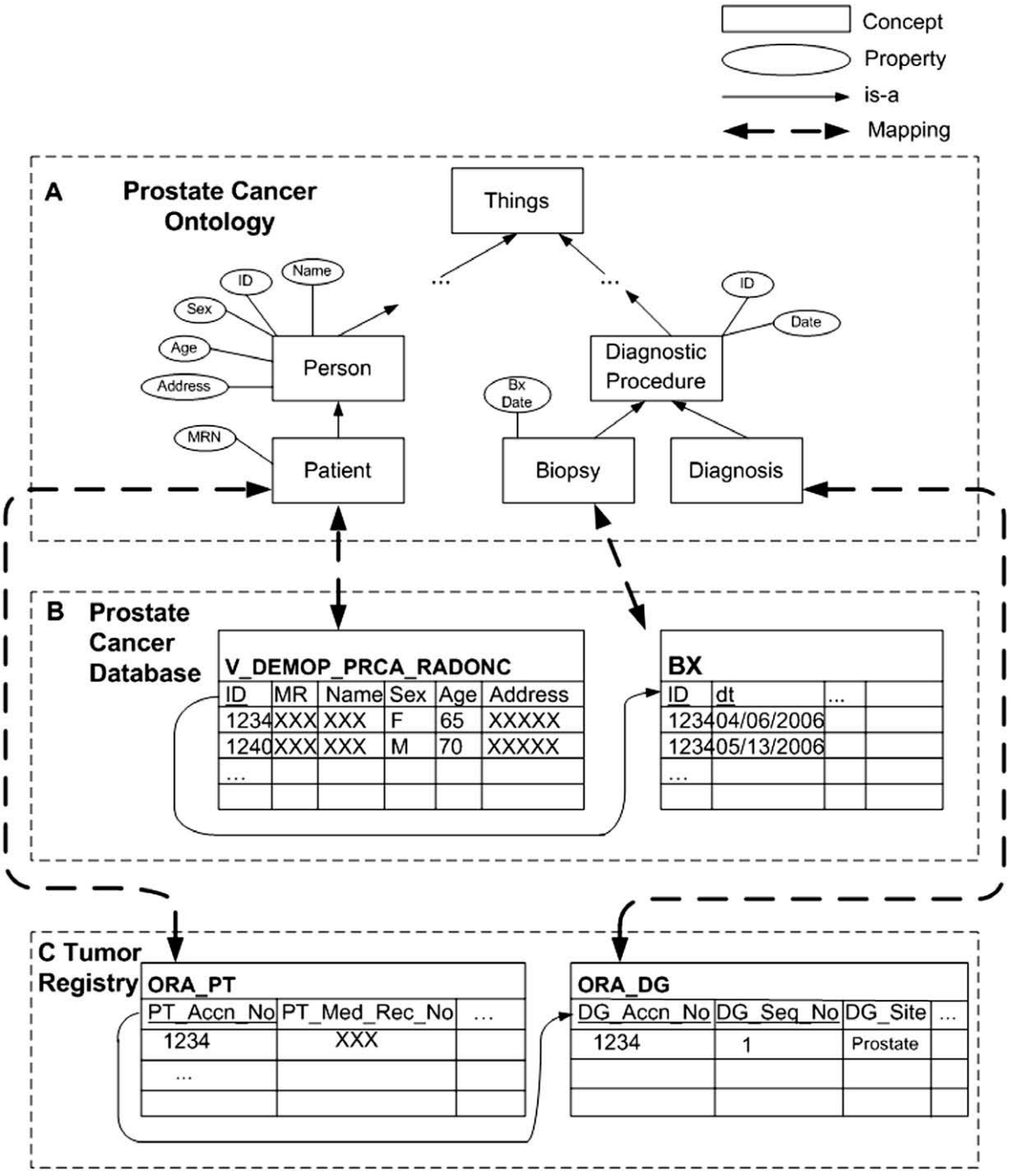

Fig. 2. Mappings between ontology and database schemas.

erties of the PCO concept. For example, column MR is mapped to the MRN property of the Patient concept. The MRN is an important property in PCO. It links the data together from different database systems. The Prostate Specific Antigen (PSA) is an important biomarker for prostate cancer. It is used in early detection, diagnosis, and staging of the disease, as well as the patients following radiation therapy and/or surgery and while on chemotherapy. Therefore, it is important to know the patient's PSA value change. The patient and PSA values are stored in two tables (i.e., the V_DEMOP_PRCA_RADONC and the PSA) in the database, respectively. It is important to retrieve the patient's PSA information by joining these two tables based on the MRN. This example is shown in the third paragraph of Fig. 3.

2.5. Publishing and querying integrated data on semantic web

The D2R server is applied to publish the contents of Prostate Cancer Database and Tumor Registry on the Semantic Web. A web interface allows the users to compose the SPARQL queries and display the results of PCIS. It allows the users to retrieve the data from Prostate Cancer Database and Tumor Registry in an integrated fashion.

3. Results

The results of PCIS are presented in this section. First, we present PCO that is utilized by PCIS to integrate two database systems.

```
map:PATIENT a d2rq:ClassMap;
        d2rq:dataStorage map:database;
        d2rq:uriPattern "Patient-@@V_DEMOP_PRCA_RADONC.MR@@";
        d2rq:class pc:Patient;
        .
map:V_DEMOP_PRCA_RADONC__label a d2rq:PropertyBridge;
        d2rq:belongsToClassMap map:PATIENT;
        d2rq:property rdfs:label;
        d2rq:pattern "Patient-@@V_DEMOP_PRCA_RADONC.MR@@";
        .
map:Patient_PSA a d2rq:PropertyBridge;
        d2rq:belongsToClassMap map:PATIENT;
        d2rq:property pc:PSA;
        d2rq:refersToClassMap map:PSA;
        d2rq:join "PSA.ID=V_DEMOP_PRCA_RADONC.ID";
    .
map:V_DEMOP_PRCA_RADONC_MR a d2rq:PropertyBridge;
        d2rq:belongsToClassMap map:PATIENT;
        d2rq:property pc:MRN;
        d2rq:column "V_DEMOP_PRCA_RADONC.MR";
        .
map:V_DEMOP_PRCA_RADONC_SEX a d2rq:PropertyBridge;
        d2rq:belongsToClassMap map:PATIENT;
        d2rq:property pc:Sex;
        d2rq:column "V_DEMOP_PRCA_RADONC.SEX";
        .
map:V_DEMOP_PRCA_RADONC_RACE a d2rq:PropertyBridge;
        d2rq:belongsToClassMap map:PATIENT;
        d2rq:property pc:Race;
        d2rq:column "V_DEMOP_PRCA_RADONC.RACE";
        d2rq:datatype xsd:decimal;
        .
map:V_DEMOP_PRCA_RADONC_MARITAL a d2rq:PropertyBridge;
        d2rq:belongsToClassMap map:PATIENT;
        d2rq:property pc:Marital;
        d2rq:column "V_DEMOP_PRCA_RADONC.MARITAL";
        d2rq:datatype xsd:decimal;
        .
map:V_DEMOP_PRCA_RADONC_DOB a d2rq:PropertyBridge;
        d2rq:belongsToClassMap map:PATIENT;
        d2rq:property pc:DOB;
        d2rq:column "EMGORALC.V_DEMOP_PRCA_RADONC.DOB";
        d2rq:datatype xsd:date;
        .
```

Fig. 3. Sample of D2RQ mapping file.

Then, we present the mapping file that links the database schemas to the concepts of PCO. The problems needed to be addressed during the mapping are also discussed. Finally, we present the results of the query formation for the integrated system.

3.1. Prostate Cancer Ontology (PCO)

As described in Section 2.3, PCO was developed by merging the prostate cancer related concepts from NCI Thesaurus and FMA. The concepts of PCO are connected by the is-a relationships that form a Directed Acyclic Graph (DAG) structure. PCO contains 412 concepts that are organized in nine layers of the hierarchy. Twelve concepts are introduced from the domain knowledge or the requirements. Neither NCI Thesaurus nor FMA contains these concepts. Twenty-one concepts are inherited from the FMA. The rest of the concepts are inherited from the NCI Thesaurus. A sample hierarchical structure for PCO is presented in Fig. 4. The treatment options for prostate cancer include chemotherapy, radiation therapy, hormone therapy, watchful waiting, surgery, and cryotherapy. They are the children of the therapeutic procedure (see first level of Fig. 4).

Besides the hierarchical relationship among the concepts, the properties/attributes of the concepts are introduced as the extension of PCO. The properties are derived from the Prostate Cancer Database and Tumor Registry as the traits of the concepts. For example, the properties of the Patient contain MRN, name, age, sex, and address (see Fig. 2). The properties of the Patient, Surgical Pathology Report, and Biopsy Report of PCO are presented in Table 1. The properties provide the framework for the mappings between PCO concepts and the data in the database systems.

3.2. Mappings between ontology and database systems

The most important part in PCIS is to build up the mappings between PCO and two database systems, Prostate Cancer Database and Tumor Registry. The mapping between each database component and its corresponding ontology component is not a simple one to one mapping. It may require specifying some necessary transformations.

There are three types of the mappings: (1) One database table is mapped to one concept of the ontology. For example, the table BX in the Prostate Cancer Database is mapped to the concept Biopsy of PCO (see Fig. 2). The columns of the table BX such as biopsy date, type, and site, are mapped to the properties of the Biopsy. (2) One database table is mapped to more than one concept of the ontology. In other words, different columns may be mapped to different concepts. The Tumor Registry consists of the case abstracts of the medical records for neoplasm patients. It contains the summary reports for the treatments and the diagnosis information for the patients who are treated at FCCC. For example, the diagnosis table ORA_DG in the Tumor Registry includes clinical stage, pathologic stage, surgical margins, histology type, primary site, distant metastasis site, tumor size, tumor markers and etc. They are mapped to several concepts of PCO such as Diagnosis, Pathology Report, and Tumor Marker and so on. For example, the columns such as DG_GRADE, DG_HISTOLOGY, DG_PATH_T, DG_PATH_N, and DG_PATH_M in the table ORA_DG which contains the information regarding tumor grade, histology type, pathological stages, respectively. They are mapped to the properties of concept Pathology Report. As shown in Fig. 5, the column DG_PATH_M in the table ORA_DG is mapped to the property TNM_PATH_M of the Pathology Report. The other columns such as DG_SRG_MARGIN (Surgical Margins), DG_SRG_SUM (Surgery of Primary Site Summary), and DG_SRG_SUM_DT (Surgery Primary Site Summary Date) in the same table are mapped to another concept Surgical Pathology Report. (3) Multiple tables are mapped to one concept of the ontology. For example, the patient's surgical pathological information is stored in the SURG_PATH and the STAGE tables in the Prostate Cancer Database. The data from these tables are merged and mapped to the same concept Surgical Pathology Report.

Several semantic heterogeneity problems are identified during the integration process. One such problem is the synonym problem in which the different database systems may use different names to represent the identical meaning. For example, the name of the patient table is called V_DEMOP_PRCA_RADONC in the Prostate Cancer Database and the ORA_PT in the Tumor Registry. The synonym problem can be solved by mapping the different names to the same concept of the ontology. In this case, the two names are mapped to the same concept Patient of PCO (as shown in Fig. 2). The homonym problem is another semantic heterogeneity problem. The homonym problem is that the same name denotes different meanings in different systems. For example, there are different meanings for the date in the different tables in the two database systems. The date indicates the biopsy date in the table BX of the Prostate Cancer Database. However, the date means the diagnosis date in the table ORA_DG of the Tumor Registry. The homonym problem can be solved by mapping the term to different concepts or properties of the ontology. The date in the table BX is mapped to the Biopsy Date and the date in the table ORA_DG is mapped to the Diagnosis Date.

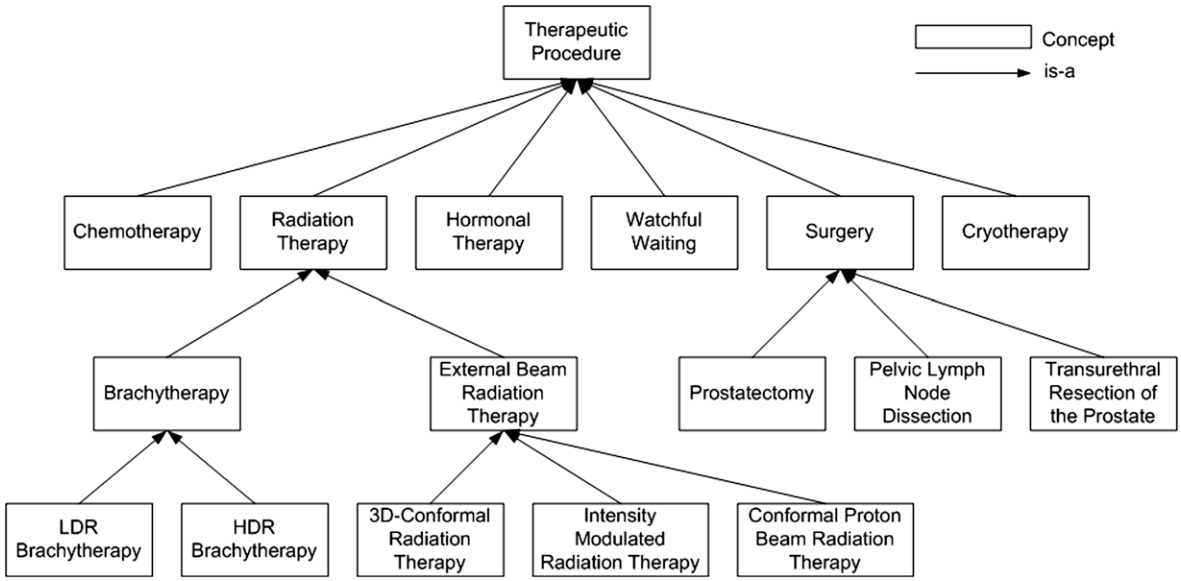

Fig. 4. Hierarchical structure for therapeutic procedure in PCO.

Table 1
Attributes for the concepts of PCO.

Concept	Attribute
Surgical pathology report	Specimen number
	Block number
	Slide number
	Highest Gleason Grade 1
	Highest Gleason Grade 2
	Highest Gleason sum
	Global Gleason Grade 1
	Global Gleason Grade 2
	Global Gleason sum
	Negative margin
	Positive margin
	Lymph node examined
	Lymph node positive
	TNM Path T
	TNM Path N
	TNM Path M
	Date
Biopsy report	Specimen number
	Block number
	Slide number
	Biopsy type
	Highest Gleason Grade 1
	Highest Gleason Grade 2
	Highest Gleason sum
	Global Gleason Grade 1
	Global Gleason Grade 2
	Global Gleason sum
	Systematic biopsy
	Old number of biopsy
	Old number positive
	TNM Path T
	TNM Path N
	TNM Path M
	Date
Patient	MRN
	Name
	Age
	Sex
	Birth country

3.3. Query formulation

In PCIS, PCO provides the conceptual level information for the data in the database systems. The mappings between the ontology and database systems enable the users to seek the low-level data fields without the detailed information of the database systems. Since PCO represents the domain knowledge of prostate cancer, the mapping also enables the users to query the data across database systems by utilizing their domain knowledge.

The web-based user interface is developed in PCIS. The web interface allows the users to compose the SPARQL queries and display the query results. The query formulation procedure is presented in the following example.

Query: Retrieve all patients in the system with tumor stage M1.

The data for this query are located in the table ORA_DG of the Tumor Registry and the table Stage of the Prostate Cancer Database (see Fig. 5). The column M in the first table and the DG_Path_M in the second table are linked to the same property TNM_Path_M of the concept Pathological Report. In order to retrieve the patients with stage M1, we need to query this property TNM_Path_M with the value "M1". The SPARQL query for this example is as follows:

```
SELECT ?patient ?stage WHERE {
?diag pc:TNM_Path_M ?stage FILTER(?stage = "M1").
?diag pc:Patient ?patient.
}
```

This query example demonstrates another semantic problem. Different tumor staging systems are used in each database system. The 1992 AJCC staging system [23] is used in the Tumor Registry. The FIGO staging system [24] is used in the Prostate Cancer Database. Since AJCC stage is applied to code the tumor stage in PCIS, an axiom is needed in the mapping file to transform the FIGO to the AJCC. The translation axiom for the tumor stage M between two coding systems is presented in Fig. 6. The mapping is performed when the users want to query the tumor staging in PCIS. Therefore, PCIS uses the translation axiom to query and retrieve the data across different database systems.

The results of the sample *Query* are presented in Fig. 7. PCIS returns the patients with the tumor stage M1 from both database systems and displays the results in the SPARQL result window (see bottom window in Fig. 7). The first record in the window comes from the Prostate Cancer Database (PCD) and the rest of records are retrieved from the Tumor Registry. In the result window, the medical record number is erased in order to block out the sensitive data of the patients.

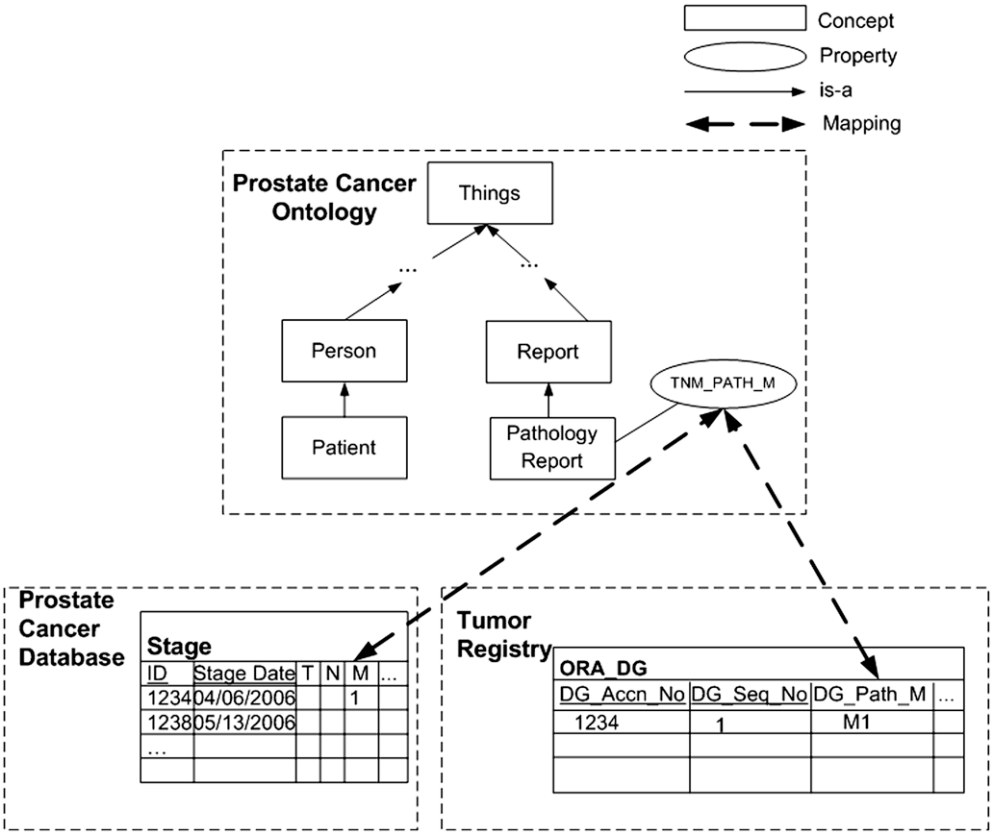

Fig. 5. A mapping example for tumor staging.

```
map:CodesTableM rdf:type d2rq:TranslationTable;
        d2rq:translation [ d2rq:databaseValue "  "; d2rq:rdfValue "Not recorded by the physician"; ];
        d2rq:translation [ d2rq:databaseValue "X "; d2rq:rdfValue "MX"; ];
        d2rq:translation [ d2rq:databaseValue "0 "; d2rq:rdfValue "M0"; ];
        d2rq:translation [ d2rq:databaseValue "1"; d2rq:rdfValue "M1"; ];
        d2rq:translation [ d2rq:databaseValue "1A"; d2rq:rdfValue "M1a"; ];
        d2rq:translation [ d2rq:databaseValue "1B"; d2rq:rdfValue "M1b"; ];
        d2rq:translation [ d2rq:databaseValue "1C"; d2rq:rdfValue "M1c"; ];
        d2rq:translation [ d2rq:databaseValue "88"; d2rq:rdfValue "Not applicable"; ];
```

Fig. 6. Translation axioms between the AJCC and the FIGO staging system.

4. Discussion

As presented in the previous section, PCO has been applied to integrate and retrieve the data from two independent database systems successfully. There are several issues that arise from the implementation of PCIS. They may be common for the ontology-based data integration too.

4.1. The properties for the concepts of PCO

The demographic information for each patient is important for prostate cancer risk factor studies. The important risk factors for prostate caner are age, ethnicity, genetic factors, and possibly dietary factors. Age is one of the most important risk factors. The incidence of prostate cancer rises rapidly after the age of 40 [40]. Prostate cancer is more common among African-American than White or Hispanic men, perhaps related to a combination of dietary and/or genetic factors [40–42]. In addition to higher incidence, the age of onset in African-American men is earlier than any other ethnic group. African-American men also have higher serum PSA levels, higher Gleason scores (signifying more pathologically aggressive disease), and more advanced stage of disease at the time of diagnosis [43].

PCIS provides the demographic information of the patients such as age, race, ethnic, religion, and birth country, etc. They are the properties of the concept Patient. Though, the concept level of ontology (i.e., NCI Thesaurus and FMA) can represent biomedical knowledge very well, it does not meet the requirements of some ontology applications (such as data integration). The properties of the concepts can enhance the capability of the ontology. They provide a framework for the ontology-based data integration and query formulation. There are 497 properties introduced for the concepts of PCO.

The properties of the concepts play an important role in the applications of ontology, and formal mechanisms are needed to define them. Two methods have been utilized to define the properties of the concepts in PCO: (1) Knowledge based: properties of this type are defined based on the understanding of the concepts. This kind of definition is intuitive and easily understandable. They are the components for the knowledge of the concepts. The prime challenge requires expertise of the concepts. For example, tumor stage, histology grade, histology type, and margins are properties

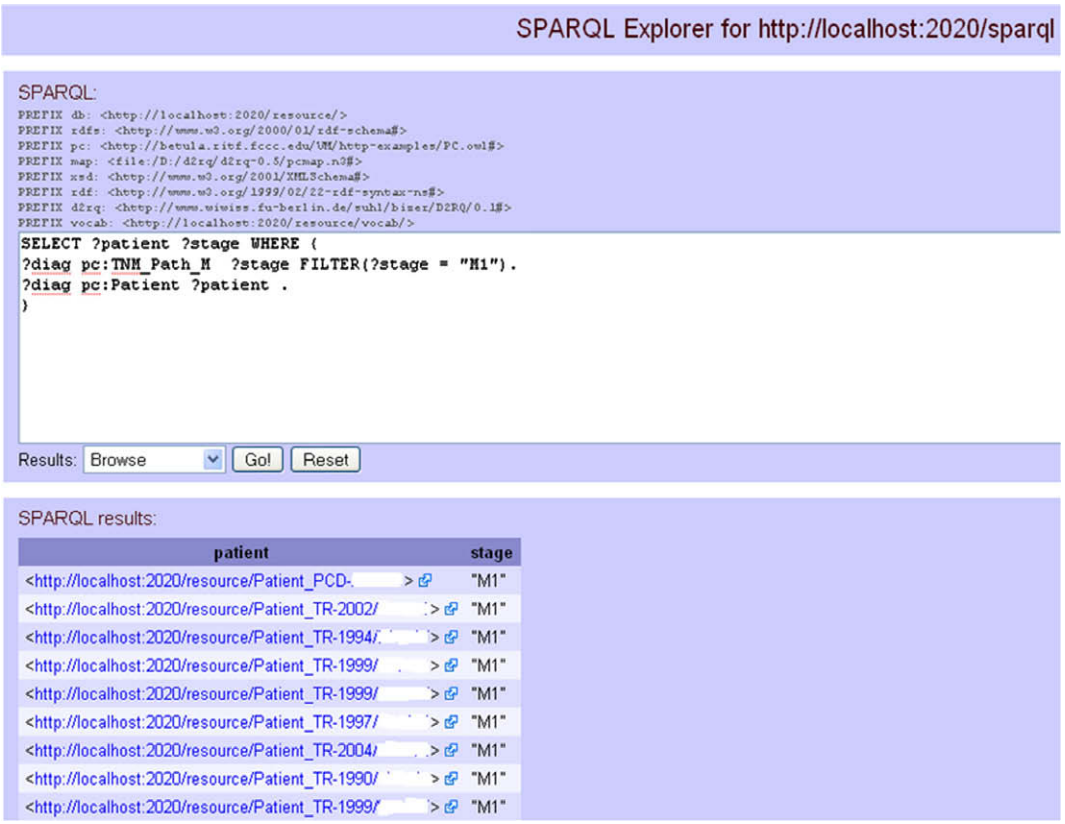

Fig. 7. Sample SPARQL query results for the prostate cancer information system.

for the Pathology Report. (2) Requirement based: some of the properties are derived from the requirements of the applications. These properties come from the integration process. For example, the date may have different meanings in tables or database systems. It is challenging to determine which concept the property belongs to. Although, this type of properties is not very sensitive to concept, the location of these properties affects the query formulation and quality of the query results.

Attributes and values have been abandoned for most ontologies in favor of a single hierarchy of qualities. Though it may emphasize the hierarchical relationships among the domain knowledge, it may not meet the requirements from the applications. Ontology constructers have already paid attention to this issue. The system of attributes and values has been built up for some ontologies. The Phenotypic Attribute Trait Ontology (PATO) is one example which is designed to be used in conjunction with ontologies of "quality-bearing entities" [44]. The Cancer Common Ontologic Representation Environment (caCORE) [45,46] utilizes a model driven architecture with concepts derived from an underlying controlled terminology provided in this case by the Enterprise Vocabulary Services. The classes or attributes of the model (i.e., UML model) are registered in a metadata repository (caDSR). caCORE provides the standard descriptions for the attributes using the registered metadata repository. caCORE enhances the semantic interoperability by encouraging its users to reuse the standard Common Data Elements (CDE). PCO merges the concepts from NCI Thesaurus and FMA. NCI Thesaurus is one of the caBIG certificated controlled terminologies. The attributes for the concepts are generated by incorporating the hierarchy structure of the domain knowledge and the requirements from the applications. It may provide a systematic way to generate and reuse the Common Data Elements. In PCIS, the database schemas are mapped to PCO which means the data model is constructed based on the ontology. caCORE, on the other hand, annotates the information models with concepts from NCI Thesaurus. PCO (concepts with their properties) serves as the information model in PCIS. It allows users to form the semantic queries utilizing concepts and properties of PCO without knowing the details of the information or data models. By contrast, caCORE requires that users have knowledge about the information models for query formulation since the ontology is not embedded in the information models. Consequently, unlike caCORE, PCIS provides ontology-directed query formulation and execution. This approach could be applied in conjunction with caCORE to enhance query services and relieve the end users of the need to have detailed knowledge of the information models of the source systems when constructing queries.

4.2. Advantages of ontology-based data integration

Although, ontology and database schemas are closely related, there are some differences between them. Ontology focuses on the representation of the domain knowledge. Database schemas focus on the data representation in a specific application. So, the wide range of the applications makes the database schemas diversified. The domain knowledge is relatively stable. The ontology is independent of any particular application, i.e., it consists of generic knowledge. The database schema has little impact on the structure of the ontology if the ontology is constructed from the domain knowledge. Ontology provides a rich, predefined vocabulary that serves as a stable conceptual interface to the database systems.

The conceptual interface provided by the ontology has the advantages of the scalability for the data integration. The individual database system is integrated into the system by mapping its contents to the corresponding components of the ontology. It is relatively straightforward to integrate the new database systems into PCIS since the conceptual framework already exists (i.e.,

PCO). It only requires updating the mapping file to include the mappings between the new database systems to PCO. For PCIS, the contents of each database system (i.e., Prostate Cancer Database or Tumor Registry) are mapped to PCO separately. It makes the integration of new database system has no impact on PCIS and the database systems integrated. Actually, either Prostate Cancer Database or Tumor Registry is still operating for its daily tasks independently. The integration of the new database systems has no impacts on both the Prostate Cancer Database and the Tumor Registry. The modifications or upgrades to either database system will not affect PCIS.

The ontology is usually developed by the domain experts. It provides a valuable resource not only for the individual system, but also for the integrated system. If the different application systems have the same domain, the same ontology can be utilized to increase the semantic interoperability among these systems.

The global query schema can be developed based on the ontology for the ontology-based data integration system. The structure of ontology makes query more intuitive for the users because it matches the domain knowledge structure. Users can formulate their queries using the concepts and properties of the ontology without the intimate knowledge of the database schemas. The mapping file indicates the location where the exact data should be retrieved because the mapping file contains the links between the database schemas and the corresponding concepts of the ontology. For example, users only need to know the TNM_Path_M property of the concept Pathology Report in PCO to retrieve all patients with cancer stage M from both database systems (as shown in the sample *Query* in Section 3.3). The properties of the concept can be found by navigating the tree structure through the ontology browser. PCIS extracts the stage information from the two database systems that are linked to TNM_Path_M in the mapping file. It also automatically performs the stage translation between the two staging systems and returns the patients from both database systems that satisfied the query criteria to the users.

Besides the prostate cancer related database systems, other domain database systems (e.g., lung cancer, breast cancer, ovarian cancer, and kidney cancer) can also be integrated into PCIS. PCO can be merged with the other domain ontologies using existing tools (e.g., Protégé PROMPT [47]). The merged ontology can provide the framework to integrate the data sources from other types of cancer.

4.3. Limitations

This work explores the ontology-based data integration methodology in prostate cancer domain. It creates a prototype system (i.e., PCIS) to integrate two database systems at FCCC. As a proof of concept, this work lacks the system evaluation. Further investigations on evaluating the system performance need to be conduced such as comparing the precision and recall between PCIS and regular database systems. This pilot project only proves that PCIS works successfully for the integration of two testing database systems. The system needs to be evaluated (e.g., system response time) when it integrates more database systems.

5. Conclusions

As a proof of concept, PCIS successfully demonstrates the procedures of the ontology-based data integration. This project shows: (1) The ontology developed from knowledge domain can be utilized for data integration. In this project, PCO is constructed by merging the concepts of prostate cancer from the NCI Thesaurus and the FMA. (2) The ontology-based data integration requires properties beyond the concept level hierarchy of the ontology. (3) The properties of the concepts can be developed based on knowledge or applications. The properties of the concepts from knowledge are intuitive. They may not meet the requirements from the applications completely so that some properties need to be introduced into the concepts from the requirements of the applications. (4) PCO serves as a stable conceptual model for data integration in PCIS. As an integrated data system, PCIS shows the advantages of the ontology-based data integration (e.g., knowledge reusing and scalability, etc.). It demonstrates the abilities to solve the semantic heterogeneity problems of the two prostate cancer related database systems, the Prostate Cancer Database and the Tumor Registry.

In the absence of the integrated data system, we need to query and retrieve the data from Prostate Cancer Database and Tumor Registry separately. PCIS provides us a unified interface to query and retrieve the data from different database systems. The ontology can be utilized to improve the user queries by guiding users to formulate their queries. The advanced ontology-based query formulation techniques will be applied to PCIS. Since PCIS integrates the Tumor Registry at FCCC, it provides a foundation to integrate data from Tumor Registry from other institutes.

Acknowledgments

This work was supported by American Cancer Society grant #IRG-92-027-14 and NIH grant P30 CA 06927.

References

[1] Lenz R. Information management in distributed healthcare networks. Data Manage Conn World 2005;3551:315–34.
[2] Brigl B et al. An integrated approach for a knowledge-based clinical workstation: architecture and experience. Method Inform Med 1998;37(1):16–25.
[3] Buccella A, Cechich A, Brisaboa NR. An ontology approach to data integration. JCS&T 2003;3(2):62–8.
[4] Wache H, et al. Ontology-based integration of information: a survey of existing approaches. In: Proceedings of the IJCAI-01 workshop: ontologies and information sharing. Seattle, WA; 2001.
[5] Kakazu KK, Cheung LW, Lynne W. The Cancer Biomedical Informatics Grid (caBIG): pioneering an expansive network of information and tools for collaborative cancer research. Hawaii Med J 2004;63(9):273–5.
[6] Fenstermacher D et al. The Cancer Biomedical Informatics Grid (caBIG™). Conf Proc IEEE Eng Med Biol Soc 2005;1:743–6.
[7] Gruber TR. A translation approach to portable ontology specifications. Knowl Acquis 1993;5(2):199–220.
[8] OWL. Web Ontology Language [cited 2009 January]. Available from: http://www.w3.org/TR/owl-features/.
[9] RDF. Resource Description Framework [cited 2009 January]. Available from: http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/.
[10] Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004;32(Database issue):D267–70.
[11] NCBO. National Center for Biomedical Ontology [cited 2009 January]. Available from: http://www.bioontology.org/.
[12] OBO. OBO [cited 2009 February]. Available from: http://obofoundry.org/.
[13] CO. Cell Ontology [cited 2009 January]. Available from: http://obo.sourceforge.net/cgi-bin/detail.cgi?cell.
[14] Whetzel PL et al. The MGED ontology: a resource for semantics-based description of microarray experiments. Bioinformatics 2006;22(7):866–73.
[15] Rosse C, Mejino Jr JL. A reference ontology for biomedical informatics: the Foundational Model of Anatomy. J Biomed Inform 2003;36(6):478–500.
[16] Golbeck J et al. The national cancer institute's thesaurus and ontology. J Web Seman 2003;1:75–80.
[17] Sioutos N et al. NCI thesaurus: a semantic model integrating cancer-related clinical and molecular information. J Biomed Inform 2007;40(1):30–43.
[18] OBI. The OBI Consortium [cited 2009 January]. Available from: http://purl.obofoundry.org/obo/obi.
[19] BFO. Basic Formal Ontology [cited 2009 January]. Available from: http://ifomis.org/bfo/.
[20] Maier A, Schnurr HP, Sure Y. Ontology-based information integration in the automotive industry. Seman Web – Iswc 2003 2003;2870:897–912.
[21] Gagnon M. Ontology-based integration of data sources, in 10th International Conference on Information Fusion 2007. Quebec, Que. p. 1–8.
[22] ACS Prostate Cancer Fact Sheets. 2008. Available from: http://www.cancer.org/downloads/PRO/ProstateCancer.pdf.
[23] Fleming ID. AJCC/TNM cancer staging, present and future. J Surg Oncol 2001;77(4):233–6.

[24] Odicino F et al. History of the FIGO cancer staging system. Int J Gynaecol Obstet 2008;101(2):205–10.
[25] Lenzerini M. Data integration: a theoretical perspective. In: PODS. 2002. p. 233–46.
[26] Ericsson R. Building business intelligence applications with .NET. 1st ed. Charles River Media; 2004.
[27] Kirov SA et al. GeneKeyDB: a lightweight, gene-centric, relational database to support data mining environments. BMC Bioinform 2005;6:72.
[28] Kemp GJL, Angelopoulos N, Gray PMD. Architecture of a mediator for a bioinformatics database federation. IEEE Trans Inform Technol Biomed 2002;6(2):116–22.
[29] Haas LM, Lin ET, Roth MA. Data integration through database federation. IBM Syst J 2002;41(4):578–96.
[30] Levy AY, Rajaraman A, Ordille JJ. Querying heterogeneous information sources using source descriptions. In: Proceedings of the twenty-second international conference on very large data bases (VLDB'96). Mumbai (Bombay), India; 1996.
[31] Baumbach J et al. CoryneRegNet: an ontology-based data warehouse of corynebacterial transcription factors and regulatory networks. BMC Genom 2006;7(1):24.
[32] Lam HY et al. Using web ontology language to integrate heterogeneous databases in the neurosciences. AMIA Annu Symp Proc 2006:464–8.
[33] D2RQ [cited 2009 January]. Available from: http://sites.wiwiss.fu-berlin.de/suhl/bizer/D2RQ/.
[34] D2R Server [cited 2009 January]. Available from: http://sites.wiwiss.fu-berlin.de/suhl/bizer/d2r-server/.
[35] SPARQL [cited 2009 January]. Available from: http://www.w3.org/TR/rdf-sparql-query/.
[36] Protégé [cited 2008 February]. Available from: http://protege.standford.edu.
[37] DataMaster [cited 2009 February]. Available from: http://protegewiki.stanford.edu/index.php/DataMaster.
[38] Barrasa J, Corcho O, Gomez-Perez A. R2O, an extensible and semantically based database-to-ontology mapping language. In: Second Workshop on Semantic Web and Databases (SWDB2004). Toronto, Canada; 2004.
[39] Laclavik M. RDB2Onto: relational database data to ontology individuals mapping. In: Navrat P, et al. editor. Tools for acquisition, organisation and presenting of information and knowledge. Slovakia: Nizke Tatry; 2006.
[40] Hankey BF et al. Cancer surveillance series: interpreting trends in prostate cancer–part I: evidence of the effects of screening in recent prostate cancer incidence, mortality, and survival rates. J Nat Cancer Inst 1999;91(12):1017–24.
[41] Reddy S et al. Prostate cancer in black and white Americans. Cancer Metast Rev 2003;22(1):83–6.
[42] Baquet CR et al. Socioeconomic factors and cancer incidence among blacks and whites. J Nat Cancer Inst 1991;83(8):551–7.
[43] Hoffman RM et al. Racial and ethnic differences in advanced-stage prostate cancer: the Prostate Cancer Outcomes Study. J Nat Cancer Inst 2001;93(5):388–95.
[44] Knowlton MN et al. A PATO-compliantzebrafish screening database (MODB): management of morpholino knockdown screen information. BMC Bioinform 2008;9:7.
[45] Komatsoulis GA et al. CaCORE version 3: implementation of a model driven, service-oriented architecture for semantic interoperability. J Biomed Inform 2008;41(1):106–23.
[46] Covitz PA et al. CaCORE: a common infrastructure for cancer informatics. Bioinformatics 2003;19(18):2404–12.
[47] Noy NF, Musen MA. The PROMPT suite: interactive tools for ontology merging and mapping. Intern J Hum-Comp Stud 2003;59(6):983–1024.