Journal of Biomedical Informatics 43 (2010) 998–1008

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Exposing the cancer genome atlas as a SPARQL endpoint

Helena F. Deus [a,b,*], Diogo F. Veiga [c], Pablo R. Freire [d], John N. Weinstein [a], Gordon B. Mills [c], Jonas S. Almeida [a]

[a] Department of Bioinformatics and Computational Biology, The University of Texas M. D. Anderson Cancer Center, 1515 Holcombe Blvd., Unit 1410, Houston, TX 77230-1402, USA
[b] Instituto de Tecnologia Química e Biológica, Universidade Nova de Lisboa, Av. da República, Estação Agronómica Nacional, 2780-157 Oeiras, Portugal
[c] Department of Systems Biology, The University of Texas M. D. Anderson Cancer Center, 7435 Fannin Street, Unit 950, Houston, TX 77030, USA
[d] Department of Molecular and Cell Biology, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA

ARTICLE INFO

Article history:
Received 7 May 2010

Keywords:
TCGA
SPARQL
RDF
Linked Data
Data integration

ABSTRACT

The Cancer Genome Atlas (TCGA) is a multidisciplinary, multi-institutional effort to characterize several types of cancer. Datasets from biomedical domains such as TCGA present a particularly challenging task for those interested in dynamically aggregating its results because the data sources are typically both heterogeneous and distributed. The Linked Data best practices offer a solution to integrate and discover data with those characteristics, namely through exposure of data as Web services supporting SPARQL, the Resource Description Framework query language. Most SPARQL endpoints, however, cannot easily be queried by data experts. Furthermore, exposing experimental data as SPARQL endpoints remains a challenging task because, in most cases, data must first be converted to Resource Description Framework triples. In line with those requirements, we have developed an infrastructure to expose clinical, demographic and molecular data elements generated by TCGA as a SPARQL endpoint by assigning elements to entities of the Simple Sloppy Semantic Database (S3DB) management model. All components of the infrastructure are available as independent Representational State Transfer (REST) Web services to encourage reusability, and a simple interface was developed to automatically assemble SPARQL queries by navigating a representation of the TCGA domain. A key feature of the proposed solution that greatly facilitates assembly of SPARQL queries is the distinction between the TCGA domain descriptors and data elements. Furthermore, the use of the S3DB management model as a mediator enables queries to both public and protected data without the need for prior submission to a single data source.

1. Introduction

The Cancer Genome Atlas (TCGA) is a multi-institutional, cross-discipline effort led by the National Cancer Institute to characterize and sequence 20 cancer types at the molecular level [1]. The results, such as the discovery of new oncogenic mutations, come with the promise of clinically relevant population stratification and have recently been widened to form a coordinated international network of similarly minded initiatives [2]. TCGA is also a valuable resource for those interested in hypothesis-driven translational research as the bulk of its data results from direct experimental evidence. The level of complexity and detail of TCGA presents both an opportunity to statistically integrate the data [3] and a challenge in its representation. Heterogeneity and distribution of data sources are characteristics almost ubiquitous in biomedical datasets, which are often made available as data services without consistent data retrieval mechanisms and formats [4]. As such, advances in translational research often require complex infrastructures to integrate data from various autonomous sources and transverse several scientific domains [5]. Even when biomedical data are exposed as Web services, these tend to reflect the heterogeneity of the data, creating a challenge for its analysis with automated tools [6]. The communities of those producing and consuming biomedical data sources have mostly agreed that wide adoption of Web services that share common protocols can greatly improve data reuse and integration without the need to locally store large quantities of data [7,8]. The Linked Data best practices [9] include a collection of standards for publishing and connecting structured data on the Web that have matured to the point of providing a practical solution for the life sciences [10], namely through use of Resource Description Framework (RDF) as a data representation formalism and SPARQL as its query language [11–13].

* Corresponding author at: Department of Bioinformatics and Computational Biology, The University of Texas M. D. Anderson Cancer Center, 1515 Holcombe Blvd., Unit 1410, Houston, TX 77230-1402, USA.

E-mail addresses: mhdeus@mdanderson.org (H.F. Deus), dveiga@mdanderson.org (D.F. Veiga), freire@bcm.edu (P.R. Freire), jweinste@mdanderson.org (J.N. Weinstein), gmills@mdanderson.org (G.B. Mills), jalmeida@mdanderson.org (J.S. Almeida).

1532-0464/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2010.09.004

1.1. Resource Description Framework

RDF is a generic model that relies on two key assertions: (a) that everything is a resource referenced by a Universal Resource Identifier (URI) and (b) that every resource is part of a triple [11]. A key feature of RDF is the separation between content and presentation, which makes it useful for transversing a variety of domains, organizations and data structures [14,15]. Datasets may be converted to RDF by identifying their data elements, which are assigned to URIs, and formalizing their relationships as triples of URIs. Common vocabularies and terminologies, such as those made available by the National Center for Biomedical Ontology [16], are often used to link different datasets. Projects such as Dbpedia [17], Bio2RDF [18], Neurocommons [19], Diseasome [20,21] and others already provide a large amount of linked biomedical data available as RDF [22].

1.2. Sparql

SPARQL, the schema-free RDF query language, was designed to allow queries to be expressed across diverse data sources based on data properties and the relationships established with other data elements rather than on the physical location of the data [23]. SPARQL queries are constructs of one or more three-element graph patterns, such as "*?Person :hasName ?Name* .", each including a subject as the first element (*?Person*), a predicate as the second element (:*hasName*) and an object as the third element (*?Name*). SPARQL graph patterns support both variable elements (for example *?Person* and *?Name*) and non-variable elements (:*hasName*), where the prefix ":" indicates the Universal Resource Locator (URL) portion of a URI. The elements specific to the domain of discourse are typically the predicates (:*hasName*), which provide an anchor for the query. The solution to a SPARQL query is a directed labeled graph reusable in future queries. These properties make SPARQL endpoints, particularly those available as Web services, a very attractive solution for biomedical data services [22] given their recurrent need for data integration methodologies and shared queries [24]. Experimental biomedical data exposed as SPARQL endpoints can greatly facilitate discovery in the life sciences as each data source can be re-used as part of query federation approaches [25].

However, several problems have been identified that hamper exposure and query of data through SPARQL endpoints without extensive technical knowledge of RDF. Notably, SPARQL is a schema-free protocol; as such formulating a query usually requires some level of eye-parsing of the data, which hinders automation [26]. Tools such as MashQL [26] or Exhibit [27] have been developed to aid in the assembly of SPARQL queries by using the underlying RDF dataset structure.

1.3. Services for an integrative infrastructure

In this report we describe an infrastructure to expose the experimental data collected by the TCGA initiative as a programmatically accessible SPARQL endpoint. TCGA experimental datasets were broken into their fundamental data elements and assigned to entities of the Simple Sloppy Semantic Database (S3DB) management model [28]. S3DB defines entities and relationships using an RDF schema (RDFS) core model that enables encapsulation of RDF triples as part of a domain description, also represented as RDF triples [29]. This solution allows both the data elements and the description of the domain to have a representation in RDF, thereby supporting SPARQL queries formulated using the domain descriptors while targeting the data elements. It is worth noting that the processing of queries in the infrastructure developed overcomes the problems associated with a static RDF representation of the data by serializing SPARQL to S3DB's protocol and query language (S3QL). A graphical tool was developed that automatically assembles SPARQL queries while navigating the description of the domain and probing the properties of its instantiation. The intended end users of the system are researchers interested in biomarker discovery that require access to both molecular raw data and clinical covariates or researchers interested in linking their own datasets to TCGA. Usage is illustrated with a case study in which biomarker identification and its biological annotation are integrated with the Diseasome dataset [20,21]. The various components of the infrastructure are made available as Representational State Transfer (REST) Web services such that each component may be re-used independently.

2. Materials and methods

The Cancer Genome Atlas is a cancer genome characterization and sequencing project generating high-throughput molecular biology data about clinical samples. That data needs to be organized, integrated and analyzed in order to identify and characterize the genomic changes in 20 cancer types. A total of 500 samples from each type of tumor were collected, along with clinical and demographic covariates. Experiments were performed by 11 distinct genomic and sequencing characterization centers (GSCCs) to obtain data regarding miRNA expression, single nucleotide polymorphisms, exon expression, DNA methylation, copy number, trace-gene-sample relationships and somatic mutations. The publicly available TCGA datasets are deposited by individual genomic characterization and sequencing centers into a shared File Transfer Protocol (FTP) location (ftp1.nci.nih.gov).

2.1. S3DB core model URIs

The S3DB engine (http://s3db.org) was used to reassemble data elements from the TCGA initiative as RDF triples. The organizational model of S3DB defines a total of seven entities that define relationships between data elements. These are: Deployment, an entity representing an instance of an S3DB engine; User, an authenticated entity of any S3DB Deployment; Project, an entity that represents a specific domain by aggregating its entities and attributes; Collection, any entity associated with a domain that may be instantiated; Rule, the association between two Collections or between a Collection and a literal attribute; Item, an instance of a Collection; Statement, the relationship between two Items or between an Item and a literal value (see Fig. 4 in [28]). By design, each instance of an S3DB entity is automatically associated with a URI that consists of a URL (identifying the S3DB deployment in which the data are kept) concatenated with an alphanumeric identifier composed of the first character of the entity name (D, U, P, C, I, R or S) and a numeric component unique for each deployment of S3DB. The TCGA domain descriptors and their relationships, i.e. the metadata describing the data, were assigned to S3DB Collections and Rules, whereas the TCGA data elements and their attributes were assigned to S3DB Items and Statements. All assignment steps were performed using the S3DB protocol (S3QL), which supports select, insert, update and delete operations.

2.2. TCGA data structures

The TCGA datasets are made available through the TCGA portal (http://cancergenome.nih.gov/) as compact assemblies of data elements with various degrees of structure: as FTP directory structures, as eXtended Markup Language (XML) and as Microarray and Gene Expression Tabular (MAGE-tab) format. MAGE-tab is a spreadsheet-based, standard format for microarray data that