Exposing The Cancer Genome Atlas (TCGA) as a SPARQL endpoint

Helena F Deus[1,2], Diogo F Veiga[1], Pablo R Freire[1], Kadir Akdemir[1], Tiago L Carvalho[1], Bradley M Broom[1], John Weistein[1] and Jonas S Almeida[1]

1The University of Texas MD Anderson Cancer Center, Department of Bioinformatics and Computational Biology, 1515 Holcombe Blvd, Texas, USA

2Instituto de Tecnologia Química e Biológica, Oeiras, Portugal

Background

• **Data, data everywhere:** Automated discovery of candidate biomarkers from multiple databases has been the central challenge in the Life Sciences in general and in the study of systemic processes such as those documented by The Cancer Genome Atlas (TCGA) in particular.

• **Foundations for a better solution**: The maturation of Semantic Web technologies offers solutions to those problems by allowing the query to be defined by navigating a formally represented domains of discourse instantiated by the data.

Objectives

We address the systems challenge of The Cancer Genome Atlas initiative (http://cancergenome.nih.gov/), which generates a large scale repository of high throughput molecular biology data generated and processed at 5 academic facilities across the USA [1, 2].

Currently, the heterogeneity of domains (genomic, transcriptomic, epigenetic effects, proteomic, clinic and demographic, etc) that are part of the TCGA data is aggregated at a single point of access charged with providing syntactic interoperability to all of the data – the TCGA portal.

The objective of this work is therefore the exposure of the highly heterogeneous TCGA data from various sources through RESTful SPARQL endpoints.

Conclusions

Using The Cancer Genome Atlas as a case study, and the S3DB (1,2,3) application as the engine of integration, we developed a data model / ontology to integrate the clinical and molecular data and expose them as **Semantic Web Services**.

Since sensitive and proprietary data is always a preoccupation with this type of studies, the core data model includes the **management of user permissions on individual data elements**.

The architecture for this novel resource provides a template for web-based solutions that bridge between data silos within a domain of knowledge and between the bench and the clinical point of care.

Methods

www.s3db.org, Nature Biotechnology (2006) 24(9):1070-1071, PLoS ONE 3(8): e2946. doi:10.1371/journal.pone.0002946

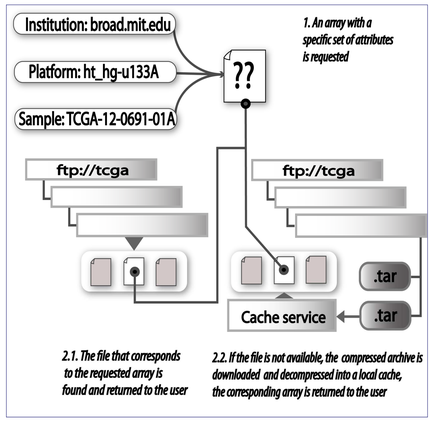


Figure 1. A caching service for TCGA archives: The TCGA datasets are typically compact assemblies of data elements with various levels of structure, of which the largest elements are the "archives" where the raw data is contained. To overcome the problem of having to bulk download each large archive, a caching service was developed to find and retrieve the latest version of an array from the TCGA archives given a set of attribute-value pairs (for institution, platform and sample (1)). The dynamic link generator, made publicly available at http://tcga.s3db.org/TCGAsync.php, recursively browses the TCGA datasets to return the raw data file corresponding to the correct array (2.1). If the archive has been compressed and the raw data file is not available as a symbolic link (2.2), then the TCGA archive is first downloaded and decompressed and the correct array is returned. It is not uncommon for an array to be requested multiple times. When the array is not being requested for the first time, the correct raw data file is downloaded from the caching service.

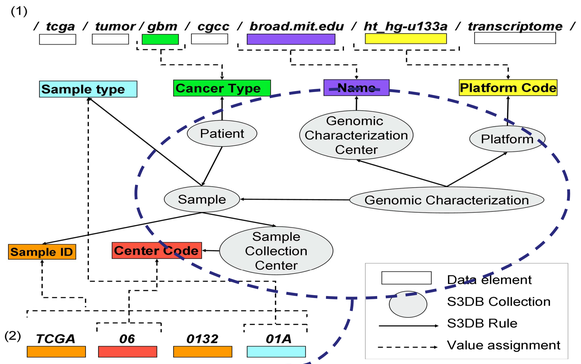
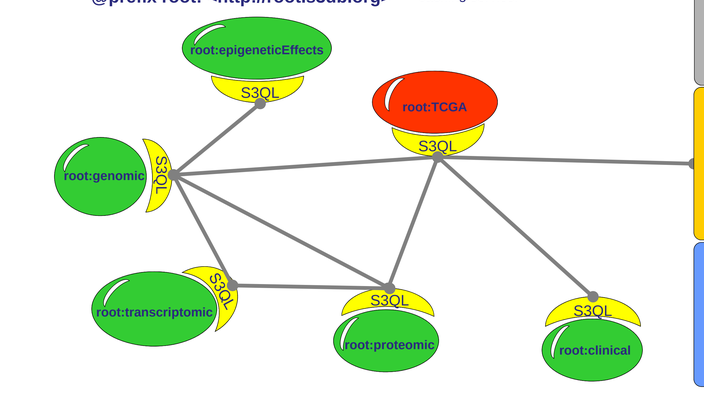
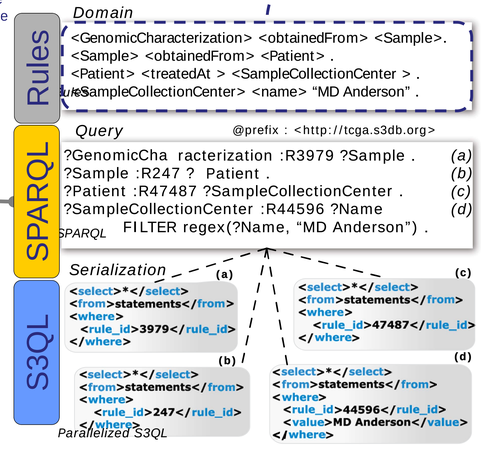


Figure 2. Identification of TCGA data elements: Breaking the TCGA datasets into their individual data elements is not a trivial task. The data elements first need to be extracted from their original hierarchical sources before they can be represented as RDF triples. For example, the path to raw data files in an FTP server (1) is separated into its constituent slash-separated portions and the resulting data elements, according to their position in the path, are assigned to values of statements. For example, the string "ht_hg-u133a" is assigned as the object of a statement where the subject is a URI identifying an instance of a platform and the predicate is the URI of the rule that links S3DB collection "Platform" and attribute "Platform Code." Similarly, each analyte is identified by a barcode (2), which broken down into its constituent elements renders the values for attributes *Sample ID, Sample Type* and *Sample Collection Center Code*, among others.





Figures 3 and 4. **Data distribution in multiple S3DB deployments by serialization of SPARQL**: each S3DB deployment (in green) is equipped with a RESTfull API (in yellow) through which data is programmatically accessed / inserted / updated using the S3DB query language (S3QL). Additionally, deployments are uniquely identified by URI which are resolved at a unique root location (http://root.s3db.org, in red). This simple system of resolving the URL of the deployments based on their unique identifiers assists the process of finding data elements which might be stored in S3DB deployments other than the one that was queried. When a query is performed where a data element that is stored in another deployment is requested, S3DB will perform the necessary S3QL query,

Privacy concerns and data sharing issues are solved by relying on the S3DB permission management model. User authentication need only occur at the deployment where the user has been registered. Once authenticated, a temporary key may be generated which is migrated with the serialized S3QL queries, providing the other deployments with the users credentials such that a level of permission may be determined.

Each S3DB deployment is also a SPARQL endpoint. Therein, SPARQL triples are serialized into S3QL queries, which are then used to query any number of deployments simultaneously. The triple patterns in the SPARQL query mimic the definition of the domain rules by replacing the predicate portion of the rule triple with its identifier.

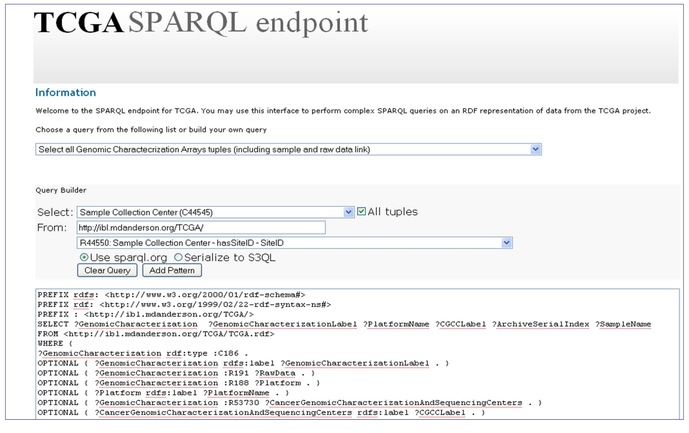


Figure 4. **An interface for TCGA SPARQL endpoint**: When the RDF representation includes a separation between the domain and its instantiation, navigating the domain to produce SPARQL triples is straightforward. This structure is imposed on the RDF representation of the data by the S3DB Core morel. Available at http://tcga.s3db.org

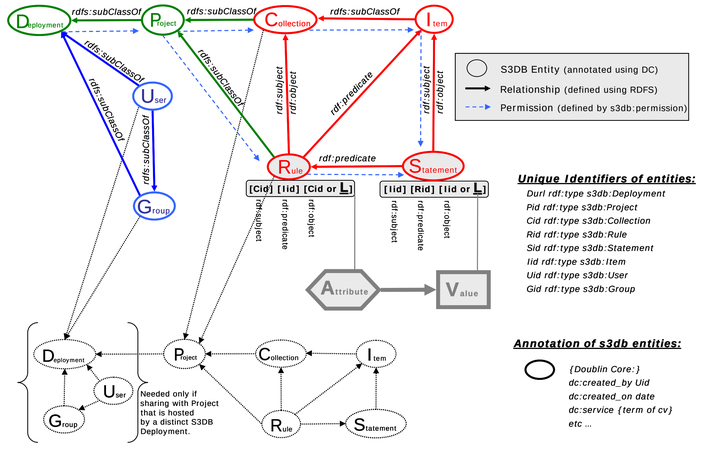


Figure 5. Permission management by embedding the corresponding propagation model in S3DB's core architecture, See Deus et. al 2008 for an illustrative application where this feature is used to "incubate" a shared data store.

References:

[1] "Comprehensive genomic characterization defines human glioblastoma genes and core pathways," *Nature,* vol. 455, pp. 1061-8, Oct 23 2008.

[2] L. Chin and J. W. Gray, "Translating insights from the cancer genome into clinical practice," *Nature,* vol. 452, pp. 553-63, Apr 3 2008.

[3] J. S. Almeida, C. Chen, R. Gorlitsky, R. Stanislaus, M. Aires-de-Sousa, P. Eleuterio, J. Carrico, A. Maretzek, A. Bohn, A. Chang, F. Zhang, R. Mitra, G. B. Mills, X. Wang, and H. F. Deus, "Data integration gets 'Sloppy'," *Nat Biotechnol,* vol. 24, pp. 1070-1, Sep 2006.

[4] H. F. Deus, R. Stanislaus, D. F. Veiga, C. Behrens, I. I. Wistuba, J. D. Minna, H. R. Garner, S. G. Swisher, J. A. Roth, A. M. Correa, B. Broom, K. Coombes, A. Chang, L. H. Vogel, and J. S. Almeida, "A Semantic Web Management Model for Integrative Biomedical Informatics," *PLoS ONE,* vol. 3, p. e2946, 2008.