Web Semantics: Science, Services and Agents on the World Wide Web 27–28 (2014) 34–41

Contents lists available at ScienceDirect

Web Semantics: Science, Services and Agents on the World Wide Web

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

Big linked cancer data: Integrating linked TCGA and PubMed

Muhammad Saleem [a,*], Maulik R. Kamdar [b], Aftab Iqbal [b], Shanmukha Sampath [b], Helena F. Deus [c], Axel-Cyrille Ngonga Ngomo [a]

[a] Universität Leipzig, IFI/AKSW, PO 100920, D-04009 Leipzig, Germany
[b] Insight Centre for Data Analytics, National University of Ireland, Galway, Ireland
[c] Foundation Medicine Inc. One Kendal Square Cambridge, MA, United States

ARTICLE INFO

Article history:
Received 15 March 2014
Received in revised form
13 June 2014
Accepted 7 July 2014
Available online 16 July 2014

Keywords:
TCGA
PubMed
RDF
Linked data
Visualization

ABSTRACT

The amount of bio-medical data available on the Web grows exponentially with time. The resulting large volume of data makes manual exploration very tedious. Moreover, the velocity at which this data changes and the variety of formats in which bio-medical data is published makes it difficult to access them in an integrated form. Finally, the lack of an integrated vocabulary makes querying this data more difficult. In this paper, we advocate the use of Linked Data to integrate, query and visualize bio-medical data. The resulting Big Linked Data allows discovering knowledge distributed across manifold sources, making it viable for the serendipitous discovery of novel knowledge. We present the concept of Big Linked Data by showing how the constant stream of new bio-medical publications can be integrated with the Linked Cancer Genome Atlas dataset (TCGA) within a virtual integration scenario. We ensure the scalability of our approach through the novel TopFed federated query engine, which we evaluate by comparing the query execution time of our system with that of FedX on Linked TCGA. Then, we show how we can harness the value hidden in the underlying integrated data by making it easier to explore through a user-friendly interface. We evaluate the usability of the interface by using the standard system usability questionnaire as well as a customized questionnaire designed for the users of our system. Our overall result of 77 suggests that our interface is easy to use and can thus lead to novel insights.

1. Introduction

Over the last years, the number of Linked Data sources published has grown to comprise more than 60 billion triples.[1] The portion of these data sources that pertain to bio-medicine is distributed across partly very large datasets. One of the newest addition to the bio-medical datasets is the Linked TCGA [1], a 20 billion-triple dataset which represents the Cancer Genome Atlas (TCGA) database in RDF. Making bio-medical datasets available as Linked Data presents the obvious advantage of easing their integration and fusion. These integrated datasets can then be used to support bio-medical experts during the analysis and exploration of bio-medical knowledge as well as for the extraction of novel knowledge from this data. Moreover, the provision of bio-medical data in RDF allows the use of the powerful query language SPARQL, which eases the selection of relevant portions of the data.

In this paper, we present a scalable framework that aims to support the serendipitous discovery of bio-medical hypotheses from Big Linked bio-medical data by providing an interface for the analysis and exploration of Big Linked Data, i.e., large volumes of Linked Data that were extracted from different sources and are updated frequently. The back-end of our application supports the management and querying of high volumes of Linked Data as well as the continuous integration of this data with other novel bio-medical data from external data streams. In this paper, we consider a subset of the Linked TCGA dataset (i.e., 10 tumors) that pertains to various cancer types and its continuous integration with the RDF data extracted from the semi-structured content of PubMed. We chose PubMed because it contains more than 23 million publications and provides an interface that allows discovering novel publications as soon as they are made available. The user interface developed on top of the resulting datasets presents an easily understandable, integrated, up-to-date view of

* Corresponding author. Tel.: +49 17666169917.
E-mail addresses: saleem.muhammd@gmail.com, saleem@informatik.uni-leipzig.de (M. Saleem), maulik.kamdar@insight-centre.org (M.R. Kamdar), aftab.iqbal@deri.org (A. Iqbal), shanmukha.sampath@deri.org (S. Sampath), hdeus@foundationmedicine.com (H.F. Deus), ngonga@informatik.uni-leipzig.de (A.-C. Ngonga Ngomo).

[1] http://stats.lod2.eu/.

http://dx.doi.org/10.1016/j.websem.2014.07.004
1570-8268/

the information available in the back-end. The intuition behind our work is that when presented with such an interface, experts can detect unexpected correlations amongst known resources. These unexpected correlations can then form the basis for a serendipitous discovery. For example, bio-medical experts that specialize on rare cancer types are empowered to easily detect the interactions between these rare cancer types and other diseases. For example, they could discover that certain cancer tends to metastasize into cancers of particular types, leading to the question of why this particular cell migration occurs. This question could then lead to the serendipitous formulation of new research questions, e.g., pertaining to the rheology of certain cancer types.

The rest of this paper is structured as follows: we first give an overview of the architecture of our approach and show how it supports volume, velocity and variability to generate novel value from large Linked Data datasets. We then present the user interface built on top of integrated datasets as well as outline various features of our interface. The evaluation section shows that our data infrastructure outperforms the state of the art in the management of large amounts of data. Moreover, it shows that our interface can be used easily. The conclusion of the paper presents future research avenues pertaining to Big Linked Data. Note that this paper is based on the work presented in [2].

2. Architecture

The architecture of our system is shown in Fig. 1 and is explained in the following subsections.

2.1. Datasets

Our framework relies on three types of datasets that are loaded into various SPARQL endpoints (explained in Section 2.2): Linked data version of TCGA, PubMed metadata in RDF and a set of mappings between these datasets. In the following, we describe each of these datasets in detail.

2.1.1. Linked TCGA

The Cancer Genome Atlas is a pilot project started in 2005 by the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI). The goal of this project is to catalog the genomic alternations found in all cancers. The TCGA public data portal[2] gives open access to the cancer patient data and enables researcher to perform and validate their analysis on real data. Currently, TCGA data portal contains 27 635 text archives for 30 different cancer diseases and 9000 patients, leading up to a total of 32.3 TB[3] of data. Each disease data is categorized into three levels: level 1 is raw data, level 2 is normalized data, and level 3 is processed data. Most of the analysis is performed on level 3 data,[4] therefore we selected level 3 data in this paper.

Exploiting such large amount of data in bio-informatics applications is a major challenge. One has to download large archives and process the relevant text files in order to collect the actual data necessary for analysis. Further, data in the archives are not biologically linked and thus require lookups on various files. To overcome these issues, a Linked Data version of the Cancer Genome Atlas is developed [1]. The main aim of this work is to publish TCGA data as Linked Data and further made it publicly available through several SPARQL endpoints. This would enable researchers to issue a query against a SPARQL endpoint and get the required chunk of data necessary for analysis. Such query processing capability saves a lot of time and encourage cancer researchers to develop real time applications on top of TCGA data. Currently, Linked TCGA contains 20.4 billion triples for 27 cancer diseases.[5] However, in this paper we have used data from 10 different cancer diseases (7.36 billion triples) of Linked TCGA. The details about data is given in Table 1.

2.1.2. TCGA disease and genes mappings

The aim of the second dataset is to establish a bridge between the structured data contained in TCGA and the constant flow of RDF data generated by analyzing PubMed (see next subsection). In order to integrate these two datasets, diseases and genes (including their synonyms) found in Linked TCGA are required to be identified in PubMed articles metadata. To achieve this goal, we extracted a list of diseases and genes from Linked TCGA. Then, we made use of the BioPortal search API[6] to obtain a list of synonyms and their corresponding URIs for every disease and gene found in Linked TCGA.

The synonyms are later used for matching against any key term identified in PubMed article's abstract by using a biomedical named entity recognition tool known as BANNER [3]. URIs corresponding to the matched keywords are then used to actually establish links between PubMed articles and corresponding diseases and gene URIs. The parsed results are stored as RDF statements dubbed TCGA disease and genes mappings. A sample TCGA disease mapping[7] for the Bladder cancer using the Bioportal search API is shown in Listing 1.

2.1.3. Integrating PubMed articles

Our third dataset consists of PubMed articles meta data. The purpose of this dataset is to keep our system up-to-date w.r.t. the current knowledge on the cancer types contained in the underlying Linked TCGA dataset. However, given that PubMed articles are not in RDF, we wrote our script which takes into a list of cancer-related keywords[8] which are used to search for PubMed articles tagged with those keywords. All articles returned as a result based on set of keywords are transformed into RDF.

For searching purposes, we use the Entrez Programming Utilities (E-utilities)[9] which acts as an API to the Entrez system of databases at the National Centre for Biotechnology Information (NCBI). The E-Utilities provides access to all major functions of Entrez, such as, text searching in databases (e.g., PubMed), and downloading records in various formats. We use the E-Utilities API in our script to search for PubMed article IDs associated with a particular keyword. We retrieve the full metadata record of article in the list of articles retrieved through the list of keywords. The resulting PubMed article information in XML format is then transformed into RDF. Moreover, we analyze the abstract of each PubMed article retrieved previously to identify any disease or gene name using BANNER [3]. If a disease or a gene name is identified, we execute a SPARQL query on our TCGA disease and gene mapping dataset (see Section 2.1.2) to find matching resources. If the match is found, then we add a triple stating the relevance of a PubMed article to a particular disease or gene term using `skos:related` predicate. An excerpt of a PubMed article related to "bladder cancer" is shown in Listing 2.

2 https://tcga-data.nci.nih.gov/tcga/tcgaDownload.jsp.

3 https://tcga-data.nci.nih.gov/datareports/statsDashboard.htm.

4 https://tcga-data.nci.nih.gov/datareports/statsDashboard.htm.

5 http://tcga.deri.ie.

6 http://data.bioontology.org/documentation.

7 http://data.bioontology.org/search?q=Bladdercancer&exactmatch=true&apikey=eb54ca23-d4a4-4b36-8652-909538a5aedd.

8 List of keywords obtained from https://tcga-data.nci.nih.gov/tcga/.

9 http://www.ncbi.nlm.nih.gov/books/NBK25501/.

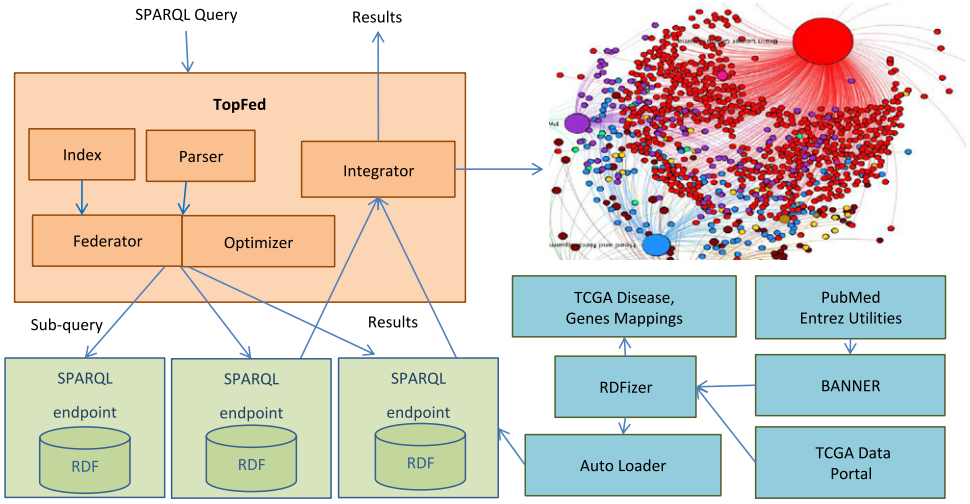

Fig. 1. Architecture of the proposed system.

Table 1
Overview of the TCGA data used.

Tumor type	Original size (GB)	Refined size (GB)	RDFized size (GB)	Triples (Million)
Cervical (CESC)	8.75	2.44	8.86	400.19
Rectal adenocarcinoma (READ)	8.07	2.25	9.04	413.31
Papillary Kidney (KIRP)	10.40	2.90	10.4	469.65
Bladder cancer (BLCA)	12.16	3.39	12.3	556.38
Acute Myeloid Leukemia (LAML)	14.85	4.14	15.1	684.05
Lower Grade Glioma (LGG)	17.08	4.76	17.1	778.82
Prostate adenocarcinoma (PRAD)	18.05	5.03	18.1	821.01
Lung squamous carcinoma (LUSC)	20.63	5.75	20.5	927.08
Cutaneous melanoma (SKCM)	23.22	6.47	23.2	1050.94
Head and neck squamous cell (HNSC)	27.6	7.69	27.5	1245.37

Listing 1: An Exemplary RDF representation of TCGA disease mappings

```
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
@prefix tcga: <http://tcga.deri.ie/schema/>.
<http://tcga.aksw.org/disease/BLCA> a tcga:disease ;
tcga:ref <http://tcga.deri.ie/graph/BLCA> ;
rdfs:label "Bladder cancer" ;
tcga:acronym "BLCA" ;
tcga:synonyms "Bladder cancer", "carcinoma of the bladder", "cancer of the bladder";
tcga:uri <http://purl.bioontology.org/ontology/MEDDRA/10005003> .
```

2.2. TopFed

After creating the datasets for our framework, we had to ensure that the user interface is responsive enough to be used in real use cases. Addressing this scalability problem is one of the major challenges when developing this platform as it relies on a very large data basis. In particular, the Linked TCGA dataset contains over 20 billion triples and is expected to reach around 30 billion triples as new data is being submitted frequently [1]. This is, to the best of our knowledge, the largest dataset of the Linked Open Data (LOD) Cloud. Hosting such large amount of data using a centralized server is simply not scalable. We thus opted for using a total of 17 SPARQL endpoints to host the 30+ billion triples which resulted from our continuous integration of PudMed articles and LinkedTCGA. Each of the endpoint contains around 2 billion of triples[10] (load balancing). Still, querying this large amount of data is a tedious problem. Therefore, we addressed this problem by developing TopFed.

TopFed is a Linked TCGA-tailored federated query engine for efficient on-the-fly data integration from multiple TCGA SPARQL endpoints. Input SPARQL query is first processed to get the individual triple patterns. For each triple pattern, the set of relevant sources are obtained using the TopFed index which contains the data distribution information among TCGA SPARQL endpoints. The annotated query is then forwarded to the federator which generates multiple sub-queries. The output of the federator is forwarded to the optimizer, which generates an optimized query execution plan for the Linked TCGA data. Each optimized sub-query is forwarded to the corresponding SPARQL endpoints and the results are integrated using the integrator. Finally, the integrated results are forwarded to the user. Complete details about TopFed can be found at the project home page.[11]

Current source selection [4–6] and SPARQL query federation [7–9] approaches are more general and do not leverage data distribution. Thus, they overestimate the set of capable sources that actually *contribute* to the final result set of SPARQL queries [6].

[10] See Section 2 for details: http://goo.gl/0oTAKV.

[11] https://code.google.com/p/topfed/.

Listing 2: An Exemplary RDF representation of meta data of a PubMed Article

```
<http://tcga.deri.ie/pubmed/22998857>
      a       <http://bio2rdf.org/pubmed_vocabulary:PubMedRecord> ;
      rdfs:label "[Urinary BLCA–4 level is useful to detect upper urinary tract urothelial cell carcinoma]." ;
      <http://bio2rdf.org/pubmed_vocabulary:author>
              <http://tcga.deri.ie/pubmed/22998857/author/4> , <http://tcga.deri.ie/pubmed/22998857/author/1> ;
      <http://bio2rdf.org/pubmed_vocabulary:chemical>
              <http://tcga.deri.ie/pubmed/22998857/chemical/4> , <http://tcga.deri.ie/pubmed/22998857/chemical/1> ;
      <http://bio2rdf.org/pubmed_vocabulary:journal>
              <http://tcga.deri.ie/pubmed/22998857/Journal> ;
      <http://bio2rdf.org/pubmed_vocabulary:mesh_heading>
              <http://tcga.deri.ie/pubmed/22998857/mesh_heading/6> , <http://tcga.deri.ie/pubmed/22998857/mesh_heading/5> ;
      <http://bio2rdf.org/pubmed_vocabulary:owner>
              "NLM" ;
      <http://bio2rdf.org/pubmed_vocabulary:publication_model>
              "Print–Electronic" ;
      <http://bio2rdf.org/pubmed_vocabulary:publication_type>
              "Journal Article" , "Research Support, Non–U.S. Gov't" , "English Abstract" ;
      <http://bio2rdf.org/pubmed_vocabulary:status>
              "MEDLINE" ;
      dcterms:abstract <http://tcga.deri.ie/pubmed/22998857/abstract> ;
      dcterms:identifier "pubmed:22998857" ;
      dcterms:language "spa" ;
      dcterms:title "[Urinary BLCA–4 level is useful to detect upper urinary tract urothelial cell carcinoma]." ;
      skos:related <http://tcga.deri.ie/graph/BLCA> .
```

An over-estimation of sources can be very expensive while dealing with Big Linked Data sources such as the integrated Linked TCGA and PubMed datasets. In contrast, TopFed is an index-assisted approach particularly designed for Linked TCGA and makes use of the intelligent data distribution provided as input.[12] Using a light-weight index, TopFed is able to detect the contributing sources for each SPARQL query and can thus reduce the number of sources selected (without losing any recall) during the query federation. By selecting fewer sources than state-of-the-art approaches, our approach can compute the answer to queries significantly faster, leading to acceptable response times for the queries required to use our framework.

3. Visualization

One of the most important challenges when delivering and using data-driven solutions for any type of human process is the provision of data visualization/summarization tools that are intuitive and easy to use for the experts [10]. Searching across integrated linked data sources, aggregating and displaying the evidence required to make informed decisions, and reusing the retrieved results to address challenges in different contexts are the main tasks of such visual analytics platforms. The current method by which physicians look for information on the Web is through peer-reviewed publications. However, with the indexing of over 10,000 papers in PubMed every year, keeping up-to-date with the literature and using this knowledge to derive new research questions has become a herculean task. To facilitate the intuitive exploration of the information available in TCGA datasets and tumor-related publications, we coupled the integrated Linked TCGA and PubMed data with an interactive visual analytics platform called the Linked TCGA Dashboard. A prototype of the interface is available at http://srvgal78.deri.ie/tcga-pubmed/.

3.1. Network explorer

The Linked TCGA Dashboard comprises different perspectives through which the underlying integrated data sources can be explored. The Network Explorer perspective features a highly dense, force-directed network graph linking the different tumor typologies analyzed in TCGA to the publication resources where more information about these tumors can be discovered. The linking/display method retrieves tumor-associated publications and presents them as a bipartite network graph initially (Fig. 3), with the nodes colored according to the tumor mappings. On the selection of a publication node, the graph gets constrained to display the Mesh Terms associated with the selected publication resource. The relevant terms are displayed as distinct nodes along with other publication nodes which reference common Mesh terms (Fig. 2(B)). The metadata of the publication (author, abstract, mesh terms, chemicals cited, etc.) is retrieved using a SPARQL query and presented in the adjoining panel (Fig. 2(C)) along with a link to the original PubMed article.

3.2. Genome Browser

The phenomena of DNA Methylation is relevant for cancer progression detection. Methylation patterns in cancer cells are known to reflect the silencing or "turning-off" of cancer protecting genes (i.e. tumor suppressor genes), thus allowing the cancer to progress. On the other hand, differential expression of exons are used to build functional profiles, which provide insights into the underlying cellular mechanisms in specific conditions. In addition to the Network Explorer perspective, the Linked TCGA Dashboard also enables cancer researchers to visualize the genomic datasets (DNA Methylation and Exon Expression) of the cancer patients against the human genome through the provision of a Genome Browser perspective.

Each researcher can select the chromosome for which he/she wishes to visualize the genomic datasets and the ideogrammatic representation of the selected chromosome is displayed in the first track (Fig. 4(A)). Ideograms are a schematic representation to depict staining patterns on a tightly-coiled chromosome. These Chromosome Bands (Ideograms) were downloaded from the Mapping and Sequencing Tracks Table in the Human Genome Assembly (GRCh37/hg19, Feb 2009), available at the UCSC Genome Browser[13] [11]. The coordinates and descriptions of the Protein-coding genes contained within this chromosome are retrieved

[12] TopFed Index: http://goo.gl/X6yz09.

[13] http://genome.ucsc.edu/.

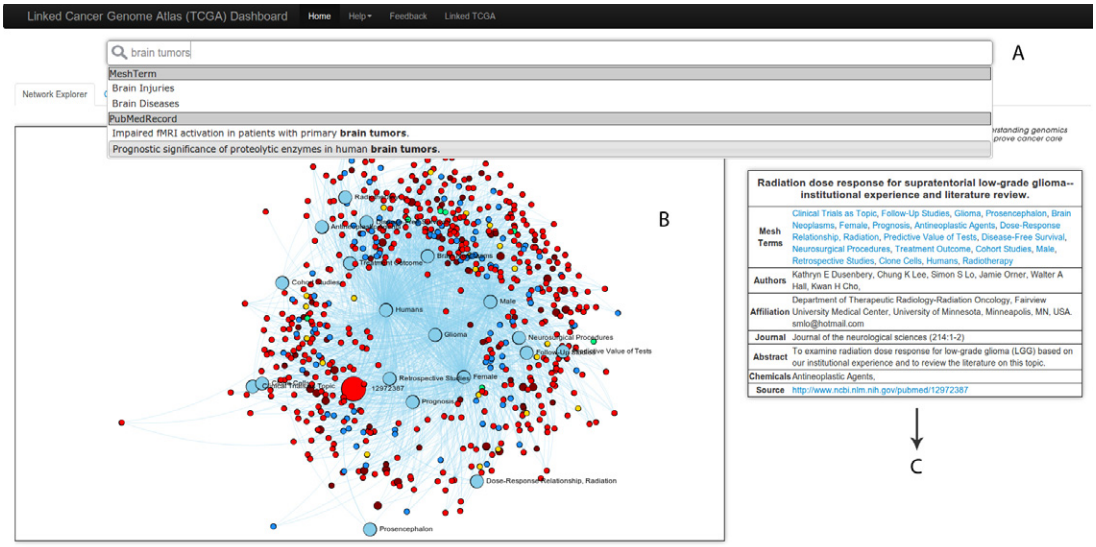

Fig. 2. Linked Cancer Genome Atlas Dashboard for the integrated visual exploration of the Linked TCGA datasets with tumor-related publications retrieved from PubMed.

from CellBase [12]. These genes are annotated using the HGNC Nomenclature [13] and the positions are indicated by start/stop attributes, and are shown in the subsequent track (Fig. 4(B)). Hovering the mouse pointer above any gene provides additional information on this gene (Fig. 4(F)).

The cancer researcher has the option to select any tumor category and load the genomic datasets of the patients diagnosed with that tumor using the interface controls (Fig. 4(E)). Selection of the patient executes SPARQL queries against the corresponding Linked TCGA endpoints and retrieves his sequencing results in real-time. These datasets (DNA Methylation and Exon Expression) are represented using bar charts (red and green respectively), whose X-coordinates are mapped to the genomic coordinates of the chromosome and the Y-coordinates indicate the normalized beta value or the RKPM value at that chromosomal position. The Genome Browser perspective allows the simultaneous comparison of these results between different patients, and the corresponding genomic datasets are stacked vertically (Fig. 4(C), (D)). The perspective also supports zooming and automatic scrolling across the length of the clicked chromosome.

3.3. Technologies

The Linked TCGA Dashboard is a Web-based client application developed using native Web technologies like HTML5 Canvas, JavaScript and JSON. The spike in the usage of HTML5 Canvas for data visualization in the recent years and its interoperability across traditional browsers, allows the application to remove the dependence on proprietary frameworks like Adobe Flash and Silverlight for interactivity.

The Linked TCGA Dashboard uses the SigmaJS[14] and the Force-Atlas graph layout algorithm [14] for the Network Explorer perspective and KineticJS[15] library, an HTML5 Canvas JavaScript framework enabling node nesting, layering, caching and event handling, for the Genome Browser perspective. The Linked TCGA Dashboard is available to download from https://github.com/maulikkamdar/tcga-pubmed/ and can be deployed using any Apache Server with PHP5 and PHP-CURL support enabled. The platform communicates with the Linked TCGA endpoints using the SPARQL 1.1 protocol and retrieves the results in JSON format.

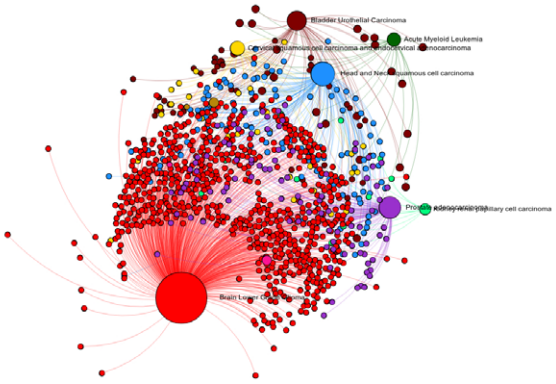

Fig. 3. Force-directed network graph linking the different TCGA tumor typologies to the publication resources. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4. Use cases

Our framework enables a variety of use cases, of which two are explained below.

4.1. Enabling Evidence-based genomic medicine

Data from TCGA is of high value for oncologists as it enables matching the evidence that they find for their own patients with those enrolled in the TCGA project, including both clinical and genomic sets. It is well known that specific genomic alterations in each individual's cancer affect response to treatment and sensitivity to drugs. As such, a physician could, for example, use our visualization to compare their own patient's methylation patterns against that for other patients enrolled in TCGA. Since genomic information in TCGA is linked with each patient's clinical prognostic and follow-up, the physician could assert, based on the similarity of genomic results, whether a patient would respond well to a given drug by observing the other patient's reaction. What this also enables is medical decisions that are highly informed by the evidence. Cancer, we now know, is a genetic disease. This means that the location where the tumor occurs (e.g. brain, liver, etc.) is less relevant for its treatment than the genetic signature that the cancer cells express (i.e. whether genes are silenced,

14 http://sigmajs.org/.
15 http://kineticjs.com/.

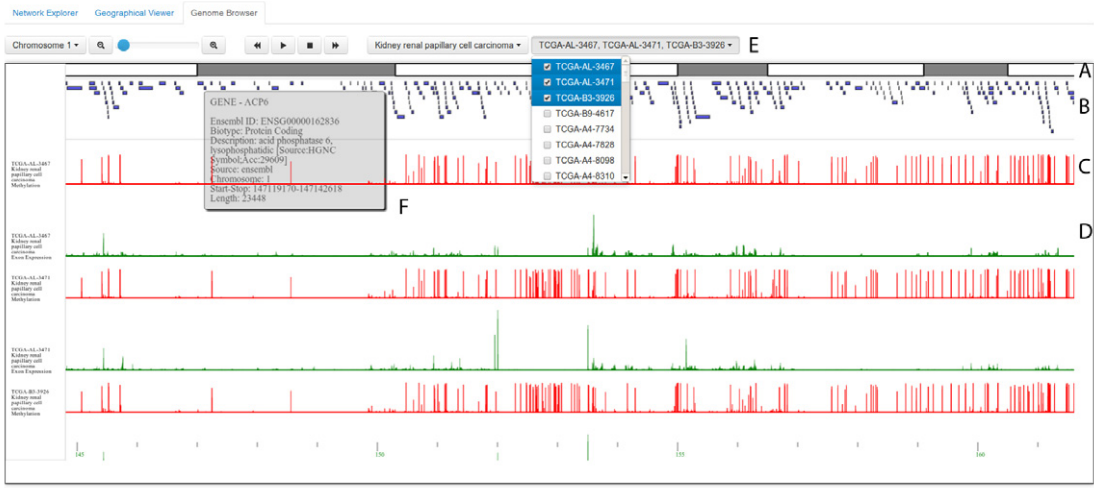

Fig. 4. Embedded Genome Browser for intuitive exploration of the Linked TCGA genomic datasets. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

amplified, etc.). However, drugs that are approved by regulatory agencies, and many publication resources, are still approved in the context of a single tumor typology. By making use of cross-resource linking, we enable the discovery on whether a drug could be applied to more than on tumor typology, by linking the two typologies through their genomic signature. As an example, a publication resource that is linked to two or more tumor typologies may mean that a discovery has been made that affects both cancer typologies and therefore the same drug or set of drugs may be applied.

4.2. Generation of new hypotheses

In addition to aiding evidence based genomic medicine, the availability of this type of linked information can also facilitate inter-disciplinary research. Some types of cancer (e.g. breast cancer) are more common than others and therefore the intricacies of their genomic signatures and genetic events tend to be more well known. However, for many rare cancers (e.g. pancreatic cancer), knowledge is more scattered and harder to find. The resource that we make available will enable researchers in the less common tumor typologies to discover association between their cancer of interest, and those that are more well studied. By finding papers where two tumor typologies co-occur, a researcher can hypothesize that the treatments and genomic events that are valid and have been proved to be relevant in the most common type of cancer, may also be relevant in the less common tumor typology. They can then exploit the genomic data in both cases to support or reject this hypothesis.

Another possible arena for hypothesis generation is that of tumor cell migration. Cancer experts have shown in the past that, for some tumor typologies, metastasis occurs preferentially in a specific tissue type. This is known as the "seed-and-soil" hypothesis, meaning that cancer cell "seeds" travelling in the blood vessels prefer some specific tissues to metastasize as they are optimal "soil" for their growth. For example, skin tumor cells preferentially metastasize in the brain. As such, co-occurrence of tumor typologies in publications may mean that cells of a particular tumor typology that is the main subject of a publication, preferentially migrate to the tissue of the second tumor typology, co-occurring in the publication but not necessarily the main subject of the paper.

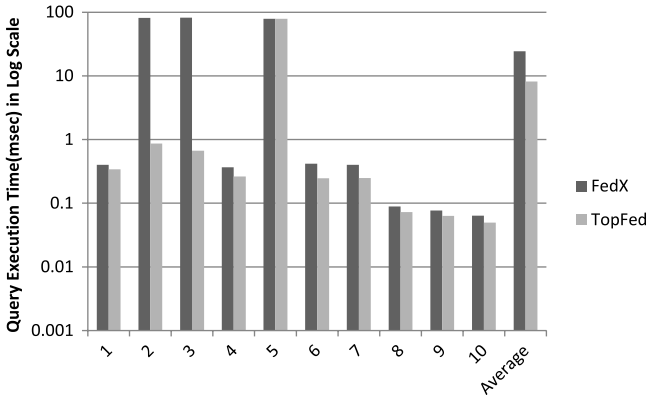

Fig. 5. Comparison of query runtimes.

5. Evaluation

The goal of our evaluation is (1) to evaluate the accuracy of Linked TCGA integration with PubMed articles, (2) to measure the performance of TopFed engine in terms of smart source selection and query execution time, and (3) to quantify the usability and usefulness of our visualization. In the following, we explain our experimental setup and the evaluation results.

5.1. Experimental setup

The aim of the TopFed evaluation is to show that it is well suited for the management of large volumes of Linked Data and can consequently support the extension of this data by novel RDF data extracted from other data streams. We thus compared TopFed with FedX [7] on 25 patients genomic results (clinical, methylation, SNP, exon-expression, gene-expression, miRNA, RNAseq2) extracted from 10 tumors. All of the data was distributed across 10 local SPARQL endpoints sharing a dedicated network. We considered 10 queries of which 4 were star-shaped [6] and the remaining queries were path-shaped or hybrid (path+star) and contained between 3 and 7 triple patterns. We ran each of the queries 10 times and present the average runtime for each of the queries. The query evaluation experiments were carried out on a 2.53 GHz i5 processor with 4 GB RAM. All of the data along with queries used for our evaluation can be found at TopFed home page.[16]

16 https://code.google.com/p/topfed/.

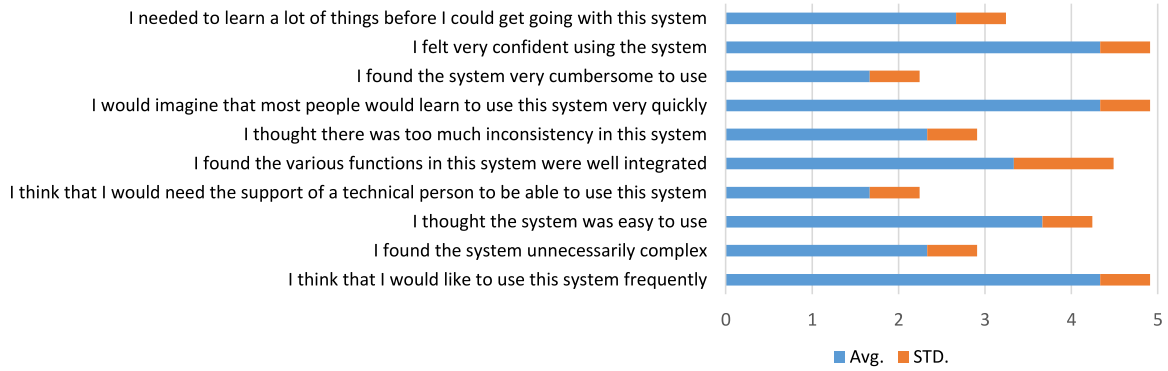

Fig. 6. Result of usability evaluation using SUS questionnaire.

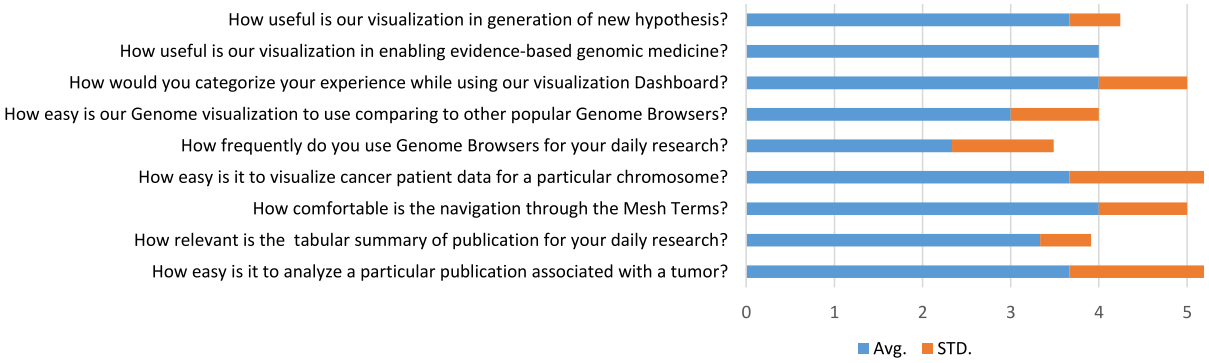

Fig. 7. Result of usefulness evaluation using our custom questionnaire.

5.2. Results

5.2.1. Data integration

In order to compute the precision of PubMed articles integration with Linked TCGA, we randomly selected 50 publications related to genes and cancer tumors from PubMed. Using our custom written scripts, we converted these articles to RDF and further evaluated the links generated using BANNER between genes and cancer diseases within those articles. For the evaluation, 3 biologists reviewed each link manually. After the manual inspection, the results of the evaluators showed that using BANNER, we have achieved a precision of **63**% for TCGA cancer disease and **84**% for TCGA genes while integrating PubMed article's metadata with Linked TCGA.

5.2.2. Query runtime performance

Fig. 5 shows the query runtime of TopFed and FedX. As an overall performance evaluation, TopFed is able to select half of the source to FedX without losing the recall. Consequently, the query run time of TopFed is about one third to that of FedX. We outperform FedX significantly on 90% of the queries. In the base case (query 2, query 3), TopFed is more than 75 times faster than FedX. We only have the same runtime for query 5. This is simply due to the number of sources selected by FedX being already optimal. Thus, our source selection approach selects exactly the same data sources. Due to smart source selection, we believe that our approach scales well on a large datasets.

5.2.3. Usability and usefulness

To assess the usability of our visualization, we used the standardized, ten-item Likert scale-based *System Usability Scale* (SUS) [15] questionnaire.[17] In addition, we conducted a custom survey[18] to asses the usefulness of our system in terms of various functionalities provided in the visualization. Both of these survey were filled by the same 3 biologists who evaluated the accuracy of PubMed articles integration with Linked TCGA. The results of SUS usability survey is shown in Fig. 6. We achieved a mean usability score of **77** indicating a high level of usability according to the SUS score. The responses to question 10 suggest that our system is adequate for frequent use (average score to question 10 $= 4.33 \pm 0.57$) by users all of type (4.33 ± 0.57 average score for question 4). The results of the usefulness of our visualization is shown in Fig. 7. As an overall usefulness evaluation, we achieved an average score of 3.52/5 indicating a usefulness of **70.37**%. In particular, we achieved an average score of 3.83/5 (76.66%) for the use cases discussed in Section 4.

6. Conclusion and future work

In this paper, we presented a scalable Linked Data-driven solution for the continuous integration of bio-medical data sources i.e., LinkedTCGA and PubMed. We also propose a visual environment that enables researchers to easily understand the data and support the serendipitous discovery of bio-medical hypotheses. The evaluation of our system leads us to believe that it is usable and useful for bio-medical experts. In future work, we aim to integrate the complete Linked TCGA data (20 billions triples) with other digital libraries. Moreover, we wish to integrate the GenomeSnip platform [16], which embodies the novel '*Genomic Wheel*' visualization for the human genome within the Linked TCGA Dashboard. By these means, we will enable cancer researchers to easily isolate genomic segments of interest and analyze the genomic datasets of the Linked TCGA patients in their context.

[17] SUS survey can be found at: http://goo.gl/kKZimO.

[18] Our custom survey can be found at: http://goo.gl/4k2fNq.

References

[1] M. Saleem, S. Shanmukha, A.-C. Ngonga Ngomo, J.S. Almeida, S. Decker, H.F. Deus, Linked cancer genome atlas database, in: I-Semantics, 2013.
[2] M. Saleem, R. Maulik, I. Aftab, S. Shanmukha, H. Deus, A.-C. Ngonga Ngomo, Fostering serendipity through big linked data, in: SWC at ISWC, 2013.
[3] R. Leaman, G. Gonzalez, BANNER: an executable survey of advances in biomedical named entity recognition, in: PSB, 2008.
[4] A. Harth, K. Hose, M. Karnstedt, A. Polleres, K.-U. Sattler, J. Umbrich, Data summaries for on-demand queries over linked data, in: WWW, 2010.
[5] M. Saleem, A.-C. Ngonga Ngomo, HiBISCuS: hypergraph-based source selection for SPARQL endpoint federation, in: ESWC, 2014.
[6] M. Saleem, A.-C. Ngonga Ngomo, J.X. Parreira, H.F. Deus, M. Hauswirth, DAW: duplicate-aware federated query processing over the Web of data, in: ISWC, 2013.
[7] A. Schwarte, P. Haase, K. Hose, R. Schenkel, M. Schmidt, FedX: optimization techniques for federated query processing on linked data, in: ISWC, 2011.
[8] O. Görlitz, S. Staab, SPLENDID: SPARQL endpoint federation exploiting void descriptions, in: COLD, 2011.
[9] M. Acosta, M.-E. Vidal, T. Lampo, J. Castillo, E. Ruckhaus, ANAPSID: an adaptive query processing engine for SPARQL endpoints, in: ISWC, 2011.
[10] M.R. Kamdar, D. Zeginis, A. Hasnain, S. Decker, H.F. Deus, ReVeaLD: a user-driven domain-specific interactive search platform for biomedical research, J. Biomed. Inform. 47 (0) (2014) 112–130.
[11] W.J. Kent, C.W. Sugnet, T.S. Furey, D. Haussler, The human genome browser at UCSC, Genome Res. (2002).
[12] M. Bleda, J. Tarraga, J. de Maria, Dopazo, I. Medina, CellBase, a comprehensive collection of RESTful Web services for retrieving relevant biological information from heterogeneous sources, Nucleic Acids Res. (2012).
[13] S. Povey, R. Lovering, E. Bruford, M. Wright, M. Lush, H. Wain, The HUGO gene nomenclature committee (HGNC), Hum. Genet. 109 (6) (2001) 678–680.
[14] M. Jacomy, S. Heymann, T. Venturini, M. Bastian, ForceAtlas2, a continuous graph layout algorithm for handy network visualization, MCR 2011.
[15] J.R. Lewis, J. Sauro, The factor structure of the system usability scale, in: HCD, 2009.
[16] M.R. Kamdar, A. Iqbal, M. Saleem, H.F. Deus, S. Decker, GenomeSnip: fragmenting the genomic wheel to augment discovery in cancer research, in: CSHALS, 2014.