A Provenance assisted Roadmap for Life Sciences Linked Open Data Cloud

Ali Hasnain[1], Qaiser Mehmood[1], Syeda Sana e Zainab[1], and Stefan Decker[1]

Insight Center for Data Analytics, National University of Ireland, Galway
firstname.lastname@insight-centre.org

Abstract. A significant portion of Web of Data is composed of multiple datasets that add high value to biomedical research. These datasets have been exposed on the web as a part of the Life Sciences Linked Open Data (LSLOD) Cloud. Different initiatives have been proposed for navigating through these datasets with or without vocabulary reuse. The significance of provenance information regarding life sciences data is great as compared to any other domain. With the provenance information, user becomes aware regarding the source, size, format along with authorization and privilege associated with the data. Previously, we proposed an approach for the creation of an active Linked Life Sciences Data Roadmap, that catalogues and links concepts as well as properties from 137 public SPARQL endpoints. In this work we extend the Roadmap with the provenance information collected directly by querying datasets. We designed a set of queries and the results were catalouged. This extended Roadmap is useful for dynamically assembling queries for retrieving data along with the provenance from multiple SPARQL endpoints. We also demonstrate its use in conjunction with other tools for selective SPARQL querying and the visualization of the LSLOD cloud. We have evaluated the performance of our approach in terms of time taken and success rates of data retrieved.

Keywords: Linked Data (LD), Provenance, SPARQL, Life Sciences (LS), Semantic Web, Query Federation

1 Introduction

A considerable portion of the Linked Open Data cloud is comprised of datasets from Life Sciences Linked Open Data (LSLOD). The significant contributors include the Bio2RDF project[1], Linked Life Data[2], Neurocommons[3], Health care and Life Sciences knowledge base[4]. This deluge of biomedical data in recent years is due to the advent of high-throughput gene sequencing technologies, that have been a primary motivation for these efforts. Although publishing datasets as RDF is a necessary step towards unified querying of biological datasets, there is

[1] http://bio2rdf.org/ (l.a.: 2014-03-31)
[2] http://linkedlifedata.com/ (l.a.: 2014-07-16)
[3] http://neurocommons.org/page/Main_Page (l.a.: 2014-07-16)
[4] http://www.w3.org/TR/hcls-kb/ (l.a.: 2014-07-16)

a critical requirement for a single interface to access the Life Sciences (LS) data. It is not sufficient to retrieve information as data being heterogeneously exposed through different public as well as private endpoints [17,2]. As LSLOD data is extremely heterogeneous and dynamic [18,8]; and integrative solutions increasingly rely on federation of queries [4]. Moreover due to the nature of domain, biologists sometimes want to get more information regarding the data including its source, creator, publisher and also statistics with respect to its size. Such information comes under the provenance spectrum and the significance of provenance information regarding life sciences data is great as compared to any other domain. To assemble queries encompassing multiple graphs distributed over different places, we previously introduced the notion of an active Roadmap for LS data – a representation of entities and the links connecting these entities from 137 public SPARQL endpoints[5], an approach described as *"a posteriori integration"*, which makes use of mapping rules that change the topology of remote graphs to match the global schema [12]. In this paper we extend our roadmap that would not only help understand which data exists in each LS SPARQL endpoint, but more importantly enable assembly of multiple source-specific federated SPARQL queries along with the provenance information available at these sources. We call it as a *Provenance Roadmap*. To generate the *Provenance Roadmap*, a set of domain independent SPARQL queries were designed to catalogue provenance information directly from the SPARQL endpoint. Our initial exploratory analysis of several LS endpoints revealed that some endpoints responded well for the designed queries whereas other lack in the direct retrieval of provenance information. Hasnain et al. [10,12] describe the methodology for developing the active Roadmap consisted of two steps: *i)* catalogue development, in which metadata is collected and analyzed, and *ii)* links creation and Roadmap development, which ensures that concepts and properties are properly mapped to a set of Query Elements (Qe) [19]. We assumed in this work that federated queries are assembled within a context that entails the initial identification of a set of Qe, in the context of cancer chemoprevention [19], identified by the domain experts participating in the EU GRANATUM project[6]. The main contribution of this publication is:

1. identification and design of initial set of queries that can contribute to retrive provenance directly from public endpoint.
2. creation of Provenance Roadmap that would lead to the provenance assisted SPARQL queries.
3. proposal of Provenance Assisted Query Engine (PAQE).
4. visual representation of Provenance Roadmap subset.
5. evaluation and analysis based on the *success rates* and *response time* for queries.

It is worth noticing that in this paper we are neither proposing a new provenance model as presented by Lebo et al. [14], nor our defined queries extensively include existing provenance vocabularies due to their limited use in the dataset

[5] `https://goo.gl/bfh7Qd`

[6] `http://www.granatum.org`(l.a.: 2015-03-05)

as mentioned by Buil–Aranda et al. [3]. The rest of this paper is organized as follows: In Section 2, we discuss the related research in line with the subject of this paper. In Section 3, we discuss the methodology and introduce the set of queries designed to collect provenance. In Section 4, we showcase some applications including a provenance assisted query engine which reasons over the Provenance Roadmap to query the LSLOD and show provenance along with the retrived data. We evaluate the results and performance in terms of time taken and query response per endpoint in Section 5.

2 Related Work

We can divide related literature into three sub-categories: SPARQL Endpoints Analysis, SPARQL Data Catalogues and Provenance for SPARQL.

SPARQL Endpoints Analysis: Buil–Aranda et al. [3] analysed 427 public SPARQL endpoints registered on the DataHub web-site [7]. It was found that:

1. VoID descriptions[8] were only available for one-in-three of the registered endpoints.
2. around one-in-six endpoints supported SPARQL 1.1 aggregate features e.g. `GROUP BY`.
3. there is reported difference between the performance of different endpoints for similar queries.
4. Service Description files for one-in-ten; that only about half the registered endpoints.
5. only one-in-three endpoints are available more than 99% of the time.
6. endpoints mostly implement result-size thresholds (10,000 being the most popular threshold; e.g. in case of virtuoso).

SPARQL Data Catalogues: Paulheim & Hertling [16] discussed the method to find a SPARQL endpoint containing content regarding any particular Linked Data URI using VoID descriptions and also uses the DataHub catalogue. Hasnain et. al [10], [12] described the process of cataloguing and linking data from LOD for Life Science domain. The proposed methodology uses catalogues that facilitate query federation. Our provenance roadmap is also a catalogue contains provenance generated by querying SPARQL endpoints.

Provenance for SPARQL: Zhao et. al [20], proposes the design patterns that represent and query provenance information relating to mapping links between heterogeneous RDF data from sources in the genomics domain. Omitola et al.[15] proposes voidp, a provenance extension for the VoID vocabulary, that allows data publishers to specify the provenance relationships of their data, while our work is focused towards querying the provenance directly from SPARQL endpoints.

[7] `http://datahub.io/` (l.a.: 2015-05-01)

[8] `http://rdfs.org/ns/void#`.

Damasio et al. [6] provides an approach capable of providing provenance information for a large fragment of SPARQL 1.1 which is based on the translation of SPARQL into relational queries over annotated relations. This works largely towards finding provenance for SPARQL queries rather than querying dataset for finding provenance. Chris et. al [1] discusses a role for provenance in quality assessment for Linked Sensor Data. Mariangiola et. al [7] proposes a notion of provenance using named graph that helps tracing where the data has been published (source) and who published it (publisher). Olaf Hartig [9] develops an automatic trust assessment approach based on provenance information. Our approach is unique from aforementioned as we focus to collect provenance directly from the data available at publicly SPARQL endpoints.

3 Methodology

Hasnain et. al [10], [12], developed an active Roadmap for navigating the LSLOD cloud and the methodology followed consists of two stages namely i) catalogue generation and ii) link generation. For Roadmap creation, data was retrieved from 137 public SPARQL endpoints[9] and organized in an RDF document - the LSLOD Catalogue. The list of SPARQL endpoints was captured from publicly available Bio2RDF datasets and by searching for datasets in CKAN[10] tagged *"life science"* or *"healthcare"* during the month of May 2014. A semi-automated method was devised to retrieve all classes (concepts) and associated properties (attributes) available through any particular endpoint by probing data instances. For collecting provenance we consider the list of endpoints already available in the Roadmap[12]. However, as noted by Buil–Aranda et al. [3], (i) many endpoints listed in the DataHub catalogue are no longer available. Hence our first step was to remove unresponsive endpoints from experimental consideration. We consider an endpoint available if it is accessible through the HTTP SPARQL protocol [5], it responds to a SPARQL-compliant query, and it returns a response in an appropriate SPARQL format; for this, we use query `Q0`: `SELECT * WHERE { ?s ?p ?o } LIMIT 1`. Issuing the query in April 2015, to 137 endpoints, 57 (41.63%)[11] responded to query `Q0`; we call these endpoints *live* and select them for collecting provenance.

3.1 Adding Provenance to the Catalogue

We designed the set of queries to collect information regarding Dataset and SPARQL endpoint directly by querying the endpoint. We call this information as Provenance. The spectrum of provenance may vary and for this research we collect basic:

- VoID statistics including totalTriples, totalDistinctClasses, totalDistinctProperties and totalDistinctObjectNodes.

[9] `http://goo.gl/ZLbLzq`

[10] `http://wiki.ckan.org/Main_Page` (l.a.: 2014-05-05)

[11] The list is available at `https://goo.gl/89PueC`.

Table 1: Dataset-level Queries - VoID statistics

№	Query
Q1	SELECT (COUNT(*) AS ?count) WHERE { ?s ?p ?o }
Q2	SELECT (COUNT(DISTINCT ?o) AS ?count) WHERE { ?s a ?o }
Q3	SELECT (COUNT(DISTINCT ?p) AS ?count) WHERE { ?s ?p ?o }
Q4	SELECT (COUNT(DISTINCT ?o) AS ?count) WHERE { ?s ?p ?o }

Table 2: Dataset-level Queries - beyond VoID statistics

№	Query
Q5	SELECT (COUNT(DISTINCT ?s) AS ?count) WHERE { ?s ?p ?o . FILTER(isBlank(?s))}
Q6	SELECT (COUNT(DISTINCT ?o) AS ?count) WHERE { ?s ?p ?o . FILTER(isLiteral(?o))}
Q7	SELECT (COUNT(DISTINCT ?o) AS ?count) WHERE { ?s ?p ?o . FILTER(isBlank(?o))}

Q8	SELECT (COUNT(DISTINCT ?b) AS ?count)
	WHERE { { ?s ?p ?b } UNION { ?b ?p ?o } FILTER(isBlank(?b)) }

– statistics beyond VoID including subjectBlankNodes, objectLiteralNodes, objectBlankNodes and totalBlankNodes
– SD information including feature, resultFormat, supportedLanguage. and dataset information including format, license, publisher, rights, retrievedOn.

For VoID statistics each query is formulated to build the equivalent of a VoID description of the dataset. We design queries to ascertain what high-level statistics, the endpoints can return about the dataset they index. We issue four queries, as listed in Table 1, to ascertain the number of triples (Q1), number of distinct classes (Q2), number of distinct properties (Q3), and number of distinct objects (Q4). These queries involve the SPARQL 1.1 features e.g. COUNT.

We further look at queries that yield statistics not supported by VoID as listed in Table 2, query (Q5- (Q8)). In particular, we experiment to see if endpoints can return a subset of statistics from the VoID Extension Vocabulary http://rdfs.org/ns/void-ext, which include counts of different types of unique RDF terms in different positions: subject blank nodes (Q5), object literal nodes (Q6), object blank nodes (Q7), all blank nodes (Q8). According to the Linked Data principles the useage of bank nodes is typically not recommended. But our exploration revealed that data publishers are still using them. It is there-

Table 3: SD and Dataset Level Provenance Queries

№	Query
Q9	SELECT distinct * WHERE { ?endpoint a sd:Service ; ?p ?o }
Q10	SELECT distinct * WHERE { ?dataset rdf:type dcterms:Dataset . ?dataset dcat:distribution ?dist . ?dist ?p ?o .} order by ?dataset ?dist ?p ?o

fore we designed a set of queries to collect information regarding balnk nodes at various positions.

Finally we issue two queries, as listed in Table 3, to ascertain the Service description of the endpoint (Q9), and dataset provenance information (Q10). For each query, we record: (i) *number of endpoints that ran the query*, and (ii) *total execution time*. Results will differ for different endpoints as we are looking at how many SPARQL endpoints responded to the queries rather than comparing performance of different endpoints. Based on the results of the queries Q1-Q10 a catalogue is generated known as Provenance Roadmap - an extract is presented in Listing 1.

3.2 An extract from Provenance Roadmap

RDFS, SD[12], PAV[13], Dublin Core[14] and VoID[15] vocabularies are used to represent the data in the Provenance Roadmap - a slice of which is presented[16]

Listing 1: An Extract from the LSLOD Catalogue for Drugbank dataset

```
<http://cu.drugbank.bio2rdf.org/sparql> a void:Dataset ;
void:class bio2rdf-drugbank_vocabulary:Drug ;
void:sparqlEndpoint <http://cu.drugbank.bio2rdf.org/sparql> ;
void:classes "141^^xsd:integer" ;
void:distinctObjects "1939582^^xsd:integer" ;
void:properties "242^^xsd:integer" ;
void:triples "4084924^^xsd:integer" ;
void-ext:distinctBlankNodes "3^^xsd:integer" ;
void-ext:distinctObjectBlankNodes "3^^xsd:integer" ;
void-ext:distinctObjectLiteralNodes "1525040^^xsd:integer" ;
void-ext:distinctSubjectBlankNodes "3^^xsd:integer" ;
dcterms:Dataset <http://identifiers.org/drugbank/>;
sd:Service <http://drugbank.bio2rdf.org/sparql#endpoint> .

bio2rdf-drugbank_vocabulary:Drug rdfs:label "Drug";
void:exampleResource <http://bio2rdf.org/drugbank:DB00313> ;
void:uriRegexPattern "^http://bio2rdf\\.org/drugbank:.*" ;
void-ext:sourceIdentifier "drugbank" .

bio2rdf-drugbank_vocabulary:patent a rdf:Property ;
rdfs:label "patent" ;
void-ext:domain bio2rdf-drugbank_vocabulary:Drug ;
void-ext:range bio2rdf-drugbank_vocabulary:Patent .

bio2rdf-drugbank_vocabulary:Patent rdfs:label "Patent" ;
void:exampleResource <http://bio2rdf.org/uspto:1338344> ;
void:uriRegexPattern "^http://bio2rdf\\.org/uspto:.*" ;
void-ext:sourceIdentifier "uspatent" .

<http://identifiers.org/drugbank/> dcat:distribution
<http://www.drugbank.ca/system/downloads/current/drugbank.xml.zip> .

<http://www.drugbank.ca/system/downloads/current/drugbank.xml.zip>
```

[12] http://www.w3.org/TR/sparql11-service-description/(l.a.: 2015-05-15)

[13] http://pav-ontology.github.io/pav/ (l.a.: 2015-05-15)

[14] http://dublincore.org/documents/dcmi-terms/ (l.a.: 2015-04-12)

[15] http://www.w3.org/TR/void/ (l.a.: 2015-04-12)

[16] In the extract, we omit URI prefixes for brevity. All prefixes can be looked up at http://prefix.cc/ (l.a.: 2015-04-31)

```
rdfs:label "DrugBank (drugbank.xml.zip)" ;
dcterms:format "application/zip" , "application/xml" ;
dcterms:license "http://www.drugbank.ca/about" ;
dcterms:publisher "http://drugbank.ca" ;
dcterms:rights "no-commercial" , "use" , "by-attribution" ;
dcterms:title "DrugBank (drugbank.xml.zip)" ;
pav:retrievedOn "2014-11-12T07:57:03-05:00^^xsd:dateTime" ;
foaf:page "http://drugbank.ca" .

<http://cu.drugbank.bio2rdf.org/sparql#endpoint>
sd:feature sd:UnionDefaultGraph , sd:DereferencesURIs ;
sd:resultFormat formats:Turtle, formats:SPARQL_Results_JSON,
formats:RDF_XML, formats:SPARQL_Results_CSV, formats:N-Triples,
formats:N3, formats:SPARQL_Results_XML, formats:RDFa;
sd:supportedLanguage  sd:SPARQL10Query .
```

Listing 1 is an illustrative example of a portion of the provenance roadmap generated for the Drugbank SPARQL endpoint[17]. VoID is used for describing the dataset: the `void#Dataset` being described in this entry is “Drugbank” SPARQL endpoint. In cases where SPARQL endpoints were available through mirrors (e.g. most Bio2RDF endpoints are available through Carleton Mirror URLs) or mentioned using alternative URLs (e.g. `http://drugbank.bio2rdf.org/sparql`), these references were also added as a second value for the `void#sparqlEndpoint` property. One identified Class(`http://bio2rdf-drugbank_vocabulary:Drug`), and one property using that class as domain (`http://bio2rdf-drugbank_vocabulary:patent`) are also included. Classes are linked to datasets using the `void#class`property; the labels were collected usually from parsing the last portion of the URI and probed instances were also recorded (`http://bio2rdf.org/drugbank:DB00313`)as values for `void#exampleResource`. Properties (`http://bio2rdf-drugbank_vocabulary:patent`) were classified as `rdfs:property`. In regard to the provenance information basic void and void-ext statistics are recorded based on results retrieved from queries listed in Table 1 and 2. This includes: `void:classes"141^^xsd:integer",void:distinctObjects"1939582^^xsd:integer", void:properties "242^^xsd:integer",void:triples "4084924^^xsd:integer", void-ext:distinctBlankNodes"3^^xsd:integer",void-ext:distinctObjectBlank Nodes "3^^xsd:integer",void-ext:distinctObjectLiteralNodes"1525040^^xsd: integer", void-ext:distinctSubjectBlankNodes "3^^xsd:integer"`. Service Description and futher provenance regarding dataset are recorded based on results retrieved from queries listed in Table 3. This includes: `sd:feature sd:UnionDefaultGraph, sd:DereferencesURIs; sd:resultFormat formats:Turtle,formats:SPARQL_Results_JSON,formats:RDF_XML,formats:N-Triples,formats:SPARQL_Results_CSV,formats:N3,formats:SPARQL_Results_XML,formats:RDFa; sd:supportedLanguage sd:SPARQL10Query` in case of service description, whereas `rdfs:label "DrugBank (drugbank.xml.zip)";dcterms:format "application/zip","application/xml" ; dcterms:license "http://www.drugbank.ca/about";dcterms:publisher "http://drugbank.ca" ; dcterms:rights "no-commercial","use","by-attribution" ; dcterms:title "DrugBank (drugbank.xml.zip)"; pav:retrievedOn "2014-11-12T07:57:03-05:00^^xsd:dateTime" ; foaf:page "http://drugbank.ca" .` in case of dataset description. It is worth

[17] `http://cu.drugbank.bio2rdf.org/sparql` (l.a.: 2015-04-01)

noticing that mentioned dataset description belongs to one distribution for one dataset whereas any SPARQL endpoint may have multiple datasets with different distributions.

4 Provenance Roadmap Applications

With the inclusion of provenance (results of the queries (Q1-10)) for the available endpoint to the existing roadmap the new Provenance Roadmap is exposed as a SPARQL endpoint[18] and relevant information is also documented[19]. As of 15th May 2014, the Roadmap consists of 263731 triples representing 1861 distinct classes and 3299 distinct properties catalogued from 137 public SPARQL endpoints. With provenence information added (May 2015), the Provenance Roadmap consists of 280064 triples representing 1861 distinct classes and 3299 distinct properties catalogued from 57 available endpoints out of 137 previously catalouged. This increase in total triple count is due to the addition of provenance information for 57 live endpoints. The remaining 80 endpoints are still part of roadmap and their correspondiing provenance information can be included in future based on their availability.

4.1 Provenance Assisted Query Engine

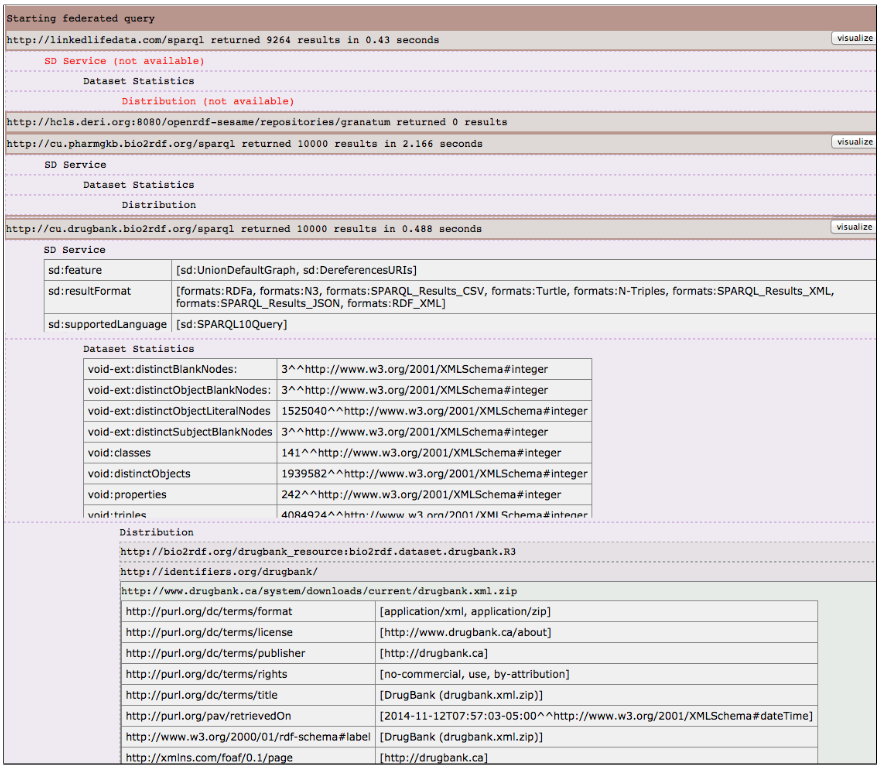
As an application of Provenance Roadmap, a Provenance Assested Query Engine (PAQE) is designed which is fundamentally a SPARQL query engine that transforms the expressions from one vocabulary into those represented using vocabularies of known SPARQL endpoints and combines those expressions into a single query using SPARQL "SERVICE" calls. The engine executes the resulting statement and returns the results to the user. The catalogued provenance organized as Provenance Roadmap is also available along with the results. DSQE is implemented on top of the Apache Jena query engine and extends it by intercepting and rewriting the SPARQL algebra.

PAQE is essentaily an extension of Domain Specifc Query Engine (DSQE) defined by Hasnain et al. [11,12] that comprises two major components: the SPARQL Algebra rewriter and the Roadmap. The algebra rewriter examines each segment of the Basic Graph Pattern (BGP) triples and attempts to expand the terms based on the vocabulary mapping into the corrosponding terms of the endpoint graphs and stores the result. Major extension to DSQE that makes it PAQE is the inclusion of provenance to the roadmap. An instance of PAQE[20] is available online that can be used to write federated SPARQL queries using drop down menu. The user can query over LSLOD using subset of predefined query elements defined in the context of drug discovery and cancer chemoprevention as mentioned earlier. For a simple query present in Listing 2, PAQE shows provenance as shown in (fig 1).

[18] `http://srvgal86.deri.ie:8000/graph/Provenance_Roadmap` (l.a.: 2015-05-15)

[19] Roadmap Homepage: `https://code.google.com/p/life-science-roadmap/`

[20] `http://srvgal86.deri.ie:8000/graph/Granatum`



```
Starting federated query
http://linkedlifedata.com/sparql returned 9264 results in 0.43 seconds
    SD Service (not available)
        Dataset Statistics
            Distribution (not available)
http://hcls.deri.org:8080/openrdf-sesame/repositories/granatum returned 0 results
http://cu.pharmgkb.bio2rdf.org/sparql returned 10000 results in 2.166 seconds
    SD Service
        Dataset Statistics
            Distribution
http://cu.drugbank.bio2rdf.org/sparql returned 10000 results in 0.488 seconds
    SD Service
```

sd:feature	[sd:UnionDefaultGraph, sd:DereferencesURIs]
sd:resultFormat	[formats:RDFa, formats:N3, formats:SPARQL_Results_CSV, formats:Turtle, formats:N-Triples, formats:SPARQL_Results_XML, formats:SPARQL_Results_JSON, formats:RDF_XML]
sd:supportedLanguage	[sd:SPARQL10Query]

```
        Dataset Statistics
```

void-ext:distinctBlankNodes:	3^^http://www.w3.org/2001/XMLSchema#integer
void-ext:distinctObjectBlankNodes:	3^^http://www.w3.org/2001/XMLSchema#integer
void-ext:distinctObjectLiteralNodes	1525040^^http://www.w3.org/2001/XMLSchema#integer
void-ext:distinctSubjectBlankNodes	3^^http://www.w3.org/2001/XMLSchema#integer
void:classes	141^^http://www.w3.org/2001/XMLSchema#integer
void:distinctObjects	1939582^^http://www.w3.org/2001/XMLSchema#integer
void:properties	242^^http://www.w3.org/2001/XMLSchema#integer
void:triples	4084924^^http://www.w3.org/2001/XMLSchema#integer

```
            Distribution
            http://bio2rdf.org/drugbank_resource:bio2rdf.dataset.drugbank.R3
            http://identifiers.org/drugbank/
            http://www.drugbank.ca/system/downloads/current/drugbank.xml.zip
```

http://purl.org/dc/terms/format	[application/xml, application/zip]
http://purl.org/dc/terms/license	[http://www.drugbank.ca/about]
http://purl.org/dc/terms/publisher	[http://drugbank.ca]
http://purl.org/dc/terms/rights	[no-commercial, use, by-attribution]
http://purl.org/dc/terms/title	[DrugBank (drugbank.xml.zip)]
http://purl.org/pav/retrievedOn	[2014-11-12T07:57:03-05:00^^http://www.w3.org/2001/XMLSchema#dateTime]
http://www.w3.org/2000/01/rdf-schema#label	[DrugBank (drugbank.xml.zip)]
http://xmlns.com/foaf/0.1/page	[http://drugbank.ca]

Fig. 1: Provenance for Drugbank Dataset

Listing 2: Sample query using PAQE Interface

```
SELECT * WHERE
{ ?x a <http://chem.deri.ie/granatum/Drug> }
```

Provencance information along with the query result help users and systems to select any particular data source based on the its size (total triple count), type of data (list of classes and properties), data distributions, the information regarding the data creator and how old was the data published. Information regarding service features, formats and supported languages can be helpful for making decisions for data itegration scenarios.

4.2 Provenance Roadmap SPARQL Service

We created Provenance Roadmap public SPARQL endpoints[21] that index the provenance information based on the results retrived for queries. Using this service, a client can query the criteria collected for a specific endpoint, or can query for endpoints matching criteria such as ordering endpoints according to number

[21] http://srvgal86.deri.ie:8000/graph/Provenance_Roadmap

of instances of a given class, or finding endpoints that use certain properties in combination with certain classes. We provide a set of sample queries to elaborate the significance of the Provenance Roadmap and the nature of questions that can be answered querying it. For example (i) Query Drugbank SPARQL Endpoint for VoID and beyond VoID statistics (Listing 3), (ii) List of Endpoints and corresponding Classes (Listing 4), (iii) List of Endpoints and corrosponding SD Services (Listing 5) and (iv) Drugbank endpoint with available datasets distribution facts (Listing 6).

Listing 3: Drugbank SPARQL Endpoint- VoID and beyond VoID statistics

```
PREFIX void: <http://rdfs.org/ns/void#>
PREFIX void-ext: <http://rdfs.org/ns/void-ext#>
SELECT distinct *
{ <http://cu.drugbank.bio2rdf.org/sparql> a void:Dataset;
void:triples ?triples;
void:classes ?classes;
void:properties ?properties;
void-ext:distinctObjectBlankNodes ?ObjectBlankNodes;
void-ext:distinctSubjectBlankNodes ?SubjectBlankNodes;
void-ext:distinctBlankNodes ?BlankNodes;
void-ext:distinctObjectLiteralNodes ?ObjectLiteralNodes. }
```

Listing 4: List of Endpoints and corresponding Classes

```
PREFIX void: <http://rdfs.org/ns/void#>
SELECT ?dataset ?class
WHERE { ?dataset void:class ?class }
order by (?dataset)
```

Listing 5: List of Endpoints and SD Services

```
PREFIX sd: <http://www.w3.org/ns/sparql-service-description#>
PREFIX void: <http://rdfs.org/ns/void#>
SELECT distinct ?endpoint ?p ?o
{ ?endpoint a void:Dataset;
sd:Service ?Service.
?Service ?p ?o. }
```

Listing 6: Drugbank endpoint and available datasets distribution facts

```
PREFIX void: <http://rdfs.org/ns/void#>
PREFIX dcterms:  <http://purl.org/dc/terms/>
PREFIX dcat: <http://www.w3.org/ns/dcat#>

SELECT distinct ?distr ?p ?o
{ <http://cu.drugbank.bio2rdf.org/sparql> a void:Dataset;
dcterms:Dataset ?dataset.
?dataset dcat:distribution ?distr.
?distr ?p ?o. }
```

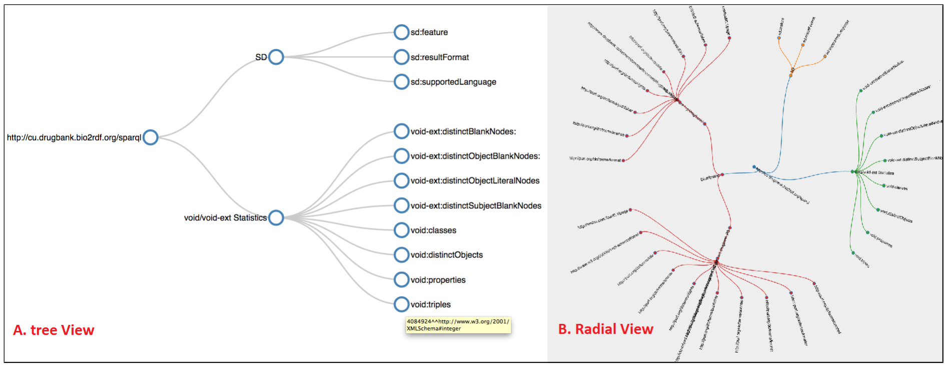

http://cu.drugbank.bio2rdf.org/sparql
SD
sd:feature
sd:resultFormat
sd:supportedLanguage
void/void-ext Statistics
void-ext:distinctBlankNodes:
void-ext:distinctObjectBlankNodes:
void-ext:distinctObjectLiteralNodes
void-ext:distinctSubjectBlankNodes
void:classes
void:distinctObjects
void:properties
void:triples
4084924^^http://www.w3.org/2001/XMLSchema#integer
A. tree View
B. Radial View

Fig. 2: Visualizing the subset of Provenance Roadmap. Tree View (A), Radial View (B)

4.3 Visualizing the Provenance Roadmap

Earlier the visualization of complete roadmap was displayed as a force-directed concept map representation in [13]. A Visualization of the provenance information is also developed to enable the domain users to visually and intuitively navigate the subset of provenance roadmap retrieved along any query result. The endpoint, its SD represenation, its VoID statistics and dataset information is visually displayed as soon as any query is executed through PAQE and user can visually explore the provenance data. Hovering over any particular node displays its information and provides value associated to that node (fig 2).

5 Results

We previously evaluated the performance of Link Generation methodology [10] by comparing it against the popular linking approaches whereas catalogues generation methodology was evaluated by analyzing the times taken to probe instances through endpoint analysis of 12 different endpoints whose underlying data sources were considered relevant for drug discovery [12]. Since Provenance Roadmap was developed by sending queries directly to the live endpoints, we evaluate and analyze our methodology in terms of comparing the success rate and response time for different queries per endpoint.

5.1 Experimental Setup

As mentioned earlier we collected a list of 137 Life Sciences public SPARQL endpoints catalogued in the roadmap in May 2014 and for collecting provenance directly from these endpoints only 57 responded to Q0, hence considered for further querying Q1 - Q10. System on which all the queries were run has a 2.53 GHz i5 processor, 8GB RAM and 320GB hard disk. For the system with Java implementation, we used Eclipse with default settings, i.e. Java Virtual Machine

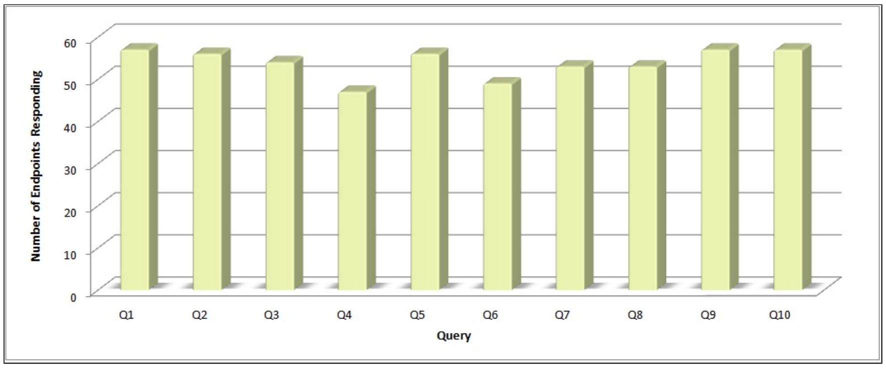


Fig. 3: SPARQL endpoints returning results per query

(JVM) initial memory allocation pool (Xms) size of 128.53MB and the maximum memory allocation pool (Xmx) size of 2057.30MB.

Success rates

We focus on the overall success rates for each query, looking at the number of endpoints that return results. The results are illustrated in Figure 3, where the success rates varying from 100% for Q1, Q9, Q10 to 82% for Q4. It is worth noticing that the queries with the highest success rates require both SPARQL 1.0, SPARQL 1.1 features to run i.e. SD-Service (Q9, Dataset Level Provenance Q10 and total number of triples (Q1)).

Response times

We also focus on runtimes for successfully executed queries, incorporating the total response time for issuing the query and streaming all results. In Figure 4, we present the runtimes for each query across all available endpoints returning non-empty results. The maximum, minimum, average and mean run time per query are ploted in the log scale. We see quite a large variance in runtimes, which is to be expected given that different endpoints host datasets of a variety of sizes and schemata on servers with a variety of computational capacity. Q6 took the maximun run time in order to retrive the result from `http://linkedlifedata.com/sparql`, whereas Q10 took minimum time for geting data from `http://pubmed.bio2rdf.org/sparql`. Q9 appears to be the quickest in terms of retriving the data with lowest average time.

6 Discussion

In this paper we describe the concept and methodology for devising Linked Life Sciences Data Provenance Roadmap. Our methodology relies on systematically

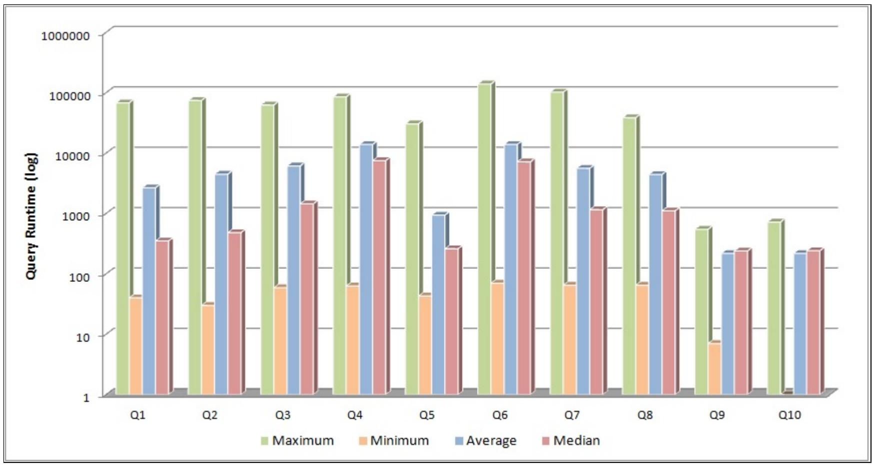


Fig. 4: Runtimes taken by different queries (Max, Min, Avg, Median)

issuing queries on various publicly available life sciences SPARQL endpoints and collecting its provenance information. We investigated the extent to which these SPARQL endpoints can be queried to collect necessary provenance information, allowing consumer agents to (e.g.) automatically select endpoints relevant to a given task. With the advent of SPARQL 1.1, new features such as aggregates allow endpoints to be interrogated for high-level statistics about the underlying dataset itself. However, it is not clear that the resulting queries would be feasible for SPARQL query engines/services to run in practice.

Investigating the issue for collecting the provenance, we proposed a set of queries (extending on existing de facto standards such as VoID, SD, DCTERM) that could be used to automatically extract provenance from a SPARQL endpoint. These queries ranged in complexity and features used. We first determined that most of the live endpoints would answer queries using novel SPARQL 1.1 features. This served as a baseline for the recall of many queries using aggregate and/or sub-query features.

We provided a list of 10 self-descriptive queries that can be issued to a SPARQL endpoints to (in theory) derive basic level of Provenence about its content that covers a large subset of VoID and beyond. The queries interrogate high-level statistics about the dataset, meta-data about its use of classes and properties, as well as information about the creater and creation date of different dataset exposed by different life science SPARQL endpoints. However, we also noted that for queries generating larger result sizes, thresholds and timeouts would likely lead to only partial results being returned.

We noticed that the number of triples per endpoint varied from few thousands to couple of hundred thousand. We also measured the frequency of usage of different URI roots (or namespaces). We found that 99.5% of class namespaces and 91% of properties namespaces are reused. This can be useful for helping

LD publishers determine the appropriate vocabulary to use. We faced multiple challenges during the roadmap development which can hinder the applicability of our approach:

- Some endpoints consider unavailable when unable to respond to simple query (`SELECT * WHERE {?s ?p ?o} LIMIT 1`).
- Some endpoints return timeout errors when a simple query (`SELECT * WHERE {?s ?p ?o} LIMIT 1`) is issued.

Nevertheless, we still found the Provenance Roadmap approach highly applicable for collecting, cataloguing and publishing provenance information after extracting directly from the live SPARQL endpoints

As future work, we would like to provide an interface for browsing the catalogue and to create a framework for updating results once a month for each new Life Sciences endpoint registered to DataHub. Though the results will indeed be partial and our evaluation has demonstrated the shortcomings, the responses we have seen for most queries are competitive. Even still, we must conclude that creating a *high quality*, *up-to-date* roadmap with *broad coverage* of all available Life Sciences public SPARQL endpoints seems infeasible: if so, this would be a notable limitation of the SPARQL infrastructure itself.

7 Conclusion

Our Provenance Roadmap is a step towards cataloguing Provenance information for LSLOD SPARQL endpoints by querying those endpoints. We also notice that static descriptions of indexed datasets in formats such as VoID remain a necessity in spite of the novel aggregation features of SPARQL 1.1. We evaluated the proposed Roadmap in terms of response time and success rate of different queries and also showcased a few applications - namely Provenance assisted Query Engine, Provenance Roadmap SPARQL Service and Provenance visualization.

Acknowledgments

This research has been supported in part by Science Foundation Ireland under Grant Number SFI/12/RC/2289.

References

1. Baillie, C., Edwards, P., Pignotti, E.: Quality assessment, provenance, and the web of linked sensor data. In: Provenance and Annotation of Data and Processes, pp. 220–222. Springer (2012)
2. Bechhofer, S., Buchan, I., De Roure, D., Missier, P., et al.: Why linked data is not enough for scientists. Future Generation Computer Systems 29(2), 599–611 (2013)
3. Buil-Aranda, C., Hogan, A., Umbrich, J., Vandenbussche, P.Y.: SPARQL Web-Querying Infrastructure: Ready for Action? In: ISWC, pp. 277–293. Springer (2013)
4. Cheung, K.H., Frost, H.R., Marshall, M.S., et al.: A journey to semantic web query federation in the life sciences. BMC bioinformatics 10(Suppl 10), S10 (2009)
5. Clark, K.G., Feigenbaum, L., Torres, E.: Sparql protocol for rdf. World Wide Web Consortium (W3C) Recommendation (2008)
6. Damásio, C.V., Analyti, A., Antoniou, G.: Provenance for sparql queries. In: The Semantic Web–ISWC 2012, pp. 625–640. Springer (2012)
7. Dezani-Ciancaglini, M., Horne, R., Sassone, V.: Tracing where and who provenance in linked data: a calculus. Theoretical Computer Science 464, 113–129 (2012)
8. Goble, C., Stevens, R., Hull, D., et al.: Data curation+ process curation= data integration+ science. Briefings in bioinformatics 9(6), 506–517 (2008)
9. Hartig, O.: Trustworthiness of data on the web. In: Proceedings of the STI Berlin & CSW PhD Workshop. Citeseer (2008)
10. Hasnain, A., Fox, R., Decker, S., Deus, H.F.: Cataloguing and linking life sciences LOD Cloud. In: 1st International Workshop on Ontology Engineering in a Data-driven World collocated with EKAW12 (2012)
11. Hasnain, A., Kamdar, M.R., Hasapis, P., Zeginis, D., Warren Jr, C.N., et al.: Linked Biomedical Dataspace: Lessons Learned integrating Data for Drug Discovery. In: International Semantic Web Conference (In-Use Track), October 2014 (2014)
12. Hasnain, A., Zainab, S.S.E., Kamdar, M.R., Mehmood, Q., Warren Jr, C., et al.: A roadmap for navigating the life scinces linked open data cloud. In: International Semantic Technology (JIST2014) conference (2014)
13. Kamdar, M.R., Zeginis, D., Hasnain, A., Decker, S., Deus, H.F.: ReVeaLD: A user-driven domain-specific interactive search platform for biomedical research. Journal of Biomedical Informatics 47(0), 112 – 130 (2014)
14. Lebo, T., Sahoo, S., McGuinness, D., Belhajjame, K., Cheney, J., Corsar, D., Garijo, D., Soiland-Reyes, S., Zednik, S., Zhao, J.: Prov-o: The prov ontology. W3C Recommendation, 30th April (2013)
15. Omitola, T., Zuo, L., Gutteridge, C., Millard, I.C., Glaser, H., Gibbins, N., Shadbolt, N.: Tracing the provenance of linked data using void. In: Proceedings of the International Conference on Web Intelligence, Mining and Semantics. p. 17. ACM (2011)
16. Paulheim, H., Hertling, S.: Discoverability of SPARQL Endpoints in Linked Open Data. In: ISWC (Posters & Demos). pp. 245–248 (2013)
17. Quackenbush, J.: Standardizing the standards. Molecular systems biology 2(1) (2006)
18. Stein, L.D.: Integrating biological databases. Nature Reviews Genetics 4(5), 337–345 (2003)
19. Zeginis, D., et al.: A collaborative methodology for developing a semantic model for interlinking Cancer Chemoprevention linked-data sources. Semantic Web (2013)
20. Zhao, J., Miles, A., Klyne, G., Shotton, D.: Linked data and provenance in biological data webs. Briefings in bioinformatics 10(2), 139–152 (2009)