Scientific Journal of Informatics
Vol. 6, No. 1, May 2019

p-ISSN 2407-7658 http://journal.unnes.ac.id/nju/index.php/sji e-ISSN 2460-0040

Dataset Characteristics Identification for Federated SPARQL Query

Nur Aini Rakhmawati[1], Lutfi Nur Fadzilah[2]

[1,2]Information System Department, Sepuluh Nopember Institute of Technology, Indonesia
Email: [1]nur.aini@is.its.ac.id, [2]lutfinf@gmail.com

Abstract

Nowadays, the amount of data published in the RDF format is increasing. Federated SPARQL query engines that are able to query from multiple distributed SPARQL endpoints have been developed recently. A federated query engine usually has different performance compared to the others. One of the factors that affect the performance of the query engine is the characteristic of the accessed RDF dataset. The aim of this work is to identify the characteristic of RDF dataset and create a query set for evaluating a federated engine. The study was conducted by identifying 16 datasets that used by 10 research papers in Linked Data area. The metrics used in this work are the number of classes, properties, entities, objects, and subjects as well as the distribution of classes and properties.

Keywords: Federated SPARQL Query, Dataset, Benchmark

1. INTRODUCTION

The discovery of Semantic Web has made the exchange of data on the web easier. Semantic Web has spawned a new standard that makes data on the web scattered around the world more structured and connected. Semantic Web also makes the data understandable by the machine to enable the use of data by applications.

The standard data exchange framework used on Semantic Web is the RDF (Resource Description Framework) [1]. Today many online databases have been created in RDF format, including DBpedia and Freebase. In order to obtain RDF data from these sources, queries should be performed using a specific query language for RDF, SPARQL (SPARQL Protocol and RDF Query Language) [2]. A web services that allows RDF data queries via SPARQL are called SPARQL endpoints.

The standard data exchange framework used on Semantic Web is the RDF (Resource Description Framework)[1]. Today many online databases have been created in RDF format, including DBpedia and Freebase. In order to obtain RDF data from these sources, queries should be performed using a specific query language for RDF, SPARQL (SPARQL Protocol and RDF Query Language) [2]. A web services that allows RDF data queries via SPARQL are called SPARQL endpoints. From year to year, the amount of data published in the

RDF format is increasing. To compensate for the rapid development of RDF data, an application system that allows RDF data queries from multiple data sources is created. RDF data queries from some SPARQL endpoints can be done using an application called a federated SPARQL query engine, namely SPLODGE [3] FedEx [4], ANAPSID [5], and ADERIS [6].

Those query engines performs different performance while using different datasets [7]. A query engine may perform well for queries on data source A, but it performs worse when it is used to query data sources B. Other query engines may work the other way around, which has a good performance for queries on data source B. Therefore, it is difficult to determine which query engine has the best performance. One of the factors causing such performance differences is the characteristics of the RDF dataset accessed by the query engine. RDF datasets certainly have different characteristics such as the number of triples, the number of classes, etc. Therefore, this work investigates characteristics of set of datasets used in federated SPARQL query. We use 16 datasets are used in 10 papers Linked Data. These datasets are identified by using a set of metrics that proposed by Duan[13] and Rakhmawati[7]. Moreover, we generate a set of queries from these datasets which can be used for federated SPARQL queries benchmark. Note that, we do not a create new benchmark suite.

The remainder of this paper is divided into six sections: Section 2 initially reviews other related works, Section 3 describes set of characteristics of datasets used in our evaluation, Section 4 explains our methodology, Section 5 explain the results of our research, and Section 5 concludes our work.

2. RELATED WORKS

Various works have been focusing on federated SPARQL queries benchmark [8-10]. Schmidt [9], proposed FedBench, a comprehensive benchmark suite for testing and analyzing the performance of federated query processing on semantic data. FedBench is highly flexible benchmark suite which is able to cover a wide range of semantic data processing strategies and use cases. Montoya [8] evaluated FedBench [9] on three federated query engines, namely ARQ1, ANAP- SID [5], and FedX [4]. The analysis has allowed to uncover hidden properties of the these systems. BioBenchmark [10] evaluated whether RDF native stores can be used to meet the needs of a biological database provider. The research evaluated five triple stores, with five biological datasets.

In terms of generating a set queries for benchmark, several works proposed how to generate queries for assessing a performance of a federated queries. Rakhmawati [11] introduced QFed, a dynamic SPARQL query set generator that takes into account the characteristics of both dataset and queries along with the cost of data communication [12]. Generated queries based logs owned by SPARQL endpoints.

Dataset is also considered in assessing federated SPARQL queries. Duan [13] proposed a metric to measure the structuredness of a dataset, which is called

coherence. The proposal of this metric is motivated by the fact that primitive data metrics, such as the number of triples and the number of literals are not enough to reveal the characteristics of the datasets.

Rakhmawati [7] investigated the relationship between the data distribution and the communication cost in a federated SPARQL query framework. A metric called Spreading Factor is proposed to compute the distribution of classes and properties on a dataset. The investigation showed that the spreading factor is correlated with the communication cost between a federated engine and the SPARQL endpoints.

In this work, we evaluate the characteristics of set of datasets used in ten papers that work on federated SPARQL queries. The metrics used for this identification are a set of metrics presented by Duan [13] and Rakhmawati [7]

3. DATASET CHARACTERISTICS

This section describes the characteristics of datasets in terms of a federated SPARQL queries.

A RDF triple consists of subject, predicate, object, while an entity is an instance of a class which its property is *rdf:type*. The following is an instance of a triple:
dbr:Indonesia rdf:type dbo:country
Where dbr:Indonesia is a subject, rdf:type is a predicate or property, dbo:country is an object, dbr:Indonesia is also an entity, and dbo:country is a class.
Suppose that, we have the following dataset from DBPEDIA which is divided into two datasets A and B:

DATASET A

```
dbr:Indonesia rdf:type dbo:country
dbr:Indonesia dbo:anthem dbr:Indonesia_Raya
dbr:Indonesia dbo:capital dbr:Jakarta
dbr:Indonesia dbo:currency dbr:Indonesian_rupiah
dbr:Malaysia rdf:type dbo:country
dbr:Malaysia dbo:anthem dbr:Negaraku
dbr:Malaysia dbo:capital dbr:Kuala_Lumpur
dbr:Malaysia dbo:currency dbr:Malaysian_ringgit
```

DATASET B

```
dbr:Philippines rdf:type dbo:country
dbr:Philippines dbo:anthem dbr:Lupang_Hinirang
dbr:Philippines dbo:capital dbr:Manila
dbr:Philippines dbo:currency dbr:Philippine_peso
```

The characteristics of the dataset above can be identified as follows:

1) Number of Subjects

There are three subjects in the dataset above, namely dbr:Indonesia, dbr:Malaysia, and dbr:Philippines.

2) Number of Properties

There are four properties in the dataset above, namely rdf:type, dbo:anthem, dbo:capital, and dbo:currency.

3) Number of Classes

There is one class in the dataset above, namely dbo:country.

4) Number of Entities

There are three entities in the dataset above, namely dbr:Indonesia, dbr:Malaysia, and dbr:Philippines.

5) Number of Triples

The number of triples shows the amount of data in a dataset that is expressed in subject-predicate-object form. There are 12 triples in the dataset above.

6) Spreading Factors (SF)

Spreading Factors is a metric used to identify the distribution of classes and properties in a dataset [13]. There are two types of Spreading Factor, namely Spreading Factor of the dataset (SF) and Spreading Factor associated with the queries (SF Q). SF Q only calculates the distribution of the classes and properties that occurs in the query.

4. METHODS

This section describes three parts, namely dataset identification, queries generation and evaluation. In a nutshell, our methodology can be presented in Figure 1. First, we collect datasets which used in papers focusing on a federated SPARQL query. Some of those datasets need to be clean up and split up before being calculated. Based on dataset calculation, queries can be generated.

.

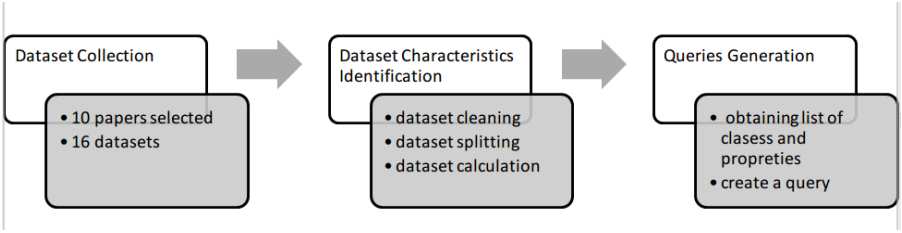

Figure 1. Methodology

2.1. Datasets

The dataset used in this study was taken from other papers which should be meeting the following criteria:

1) related to Linked Data
2) using a federated SPARQL query
3) freely assessed without permission

The papers are obtained from several sources, including Google Scholar, Science Direct, Semantic Web Journal, and International Semantic Web Conference. Some of downloaded datasets cannot be read properly by RDF data processing applications. This is generally due to an inappropriate character (bad

character) on some datasets. To solve the problem, we made the conversion of bad character to the dataset to the appropriate characters. In addition, there are also problems with the Geonames dataset where the dataset format is not recognized by the RDF data processing application. The reason is that the unusual format of RDF writing contained in the dataset. Therefore, the dataset is converted into N-triples format that can be read by RDF data processing application. The conversion process is done using script in Python [24] and RDFLib library. Table 1 explains 10 selected papers along with datasets used in the papers, while Table 2 describes the location where the datasets are accessed.

Table 1. List of Papers and Datasets Used

Number	Paper	Dataset
1	Countering language attrition with PanLex and the Web of Data	• Lexvo • DBpedia
2	How Redundant Is It? – An Empirical Analysis on Linked Datasets	• LinkedMDB • Linked Open Vocabularies (LOV) • DBLP
3	iRap - An interest-based RDF Update Propagation Framework	• DBpedia • LinkedGeoData
4	LED: curated and crowdsourced Linked Data on Music Listening Experiences	• DBpedia • British National Bibliography
5	Lexvo.org: Language Informationfor the Linguistic Linked Data Cloud	• DBpedia • YAGO • WordNet
6	Luzzu - A Framework for Linked Data Quality Assessment	• Democratic city RDF • dataset • OCD : Cidade Democrat- • ica Ontology
7	Property Path over Linked Data: Can It be Done and How to Start?	• Yago • DBpedia • Wikidata
8	Semantic Hadith: Leveraging Linked Data Opportunities for Islamic Knowledge	• QuranOntology • SemanticQuran
9	Semantic-enabled Framework for Spatial Image Information Mining of Linked Earth Observation Data	• GeoNames • DBpedia • LinkedGeoData
10	The Semantic Web Journal as Linked Data	• DBLP • Citeseer

Table 2. List of the datasets along with source and SPARQL endpoint

Dataset	Source	SPARQL endpoint
Quran Ontology	http://www.quranontology.com/	http://www.quranontology.com/Query
Semantic Quran	https://datahub.io/dataset/semanticquran	http://semanticquran.aksw.org/sparql
DBpedia	http://wiki.dbpedia.org/downloads-	https://dbpedia.org/sparql

	2016-04	
LinkedGeoData	http://downloads.linkedgeodata.org/releases/2015-11-02/	http://linkedgeodata.org/sparql
Lexvo	http://www.lexvo.org/linkeddata/resources.html	-
Yago	https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/downloads/	https://linkeddata1.calcul.u-psud.fr/sparql
WordNet	http://wordnet-rdf.princeton.edu/	http://wordnet-rdf.princeton.edu/sparql/
British National Bibliography	http://www.bl.uk/bibliographic/download.html#lodbnb	http://bnb.data.bl.uk/flint-sparql
GeoNames	http://www.geonames.org/ontology/documentation.html	-
Wikidata	https://dumps.wikimedia.org/wikidatawiki/entities/	https://query.wikidata.org/
DBLP	https://datahub.io/dataset/l3s-dblp	http://dblp.l3s.de/d2r/snorql/
Citeseer	https://datahub.io/dataset/rkb-explorer-citeseer	http://citeseer.rkbexplorer.com/sparql/
Democratic City RDF Dataset	https://datahub.io/dataset/democratic-city/resource/25bc850b-bcfb-4203-a834-f1c744a1eda7?inner_span=True	-
OCD : Cidade Democratica Ontology	http://vocab.e.gov.br/2014/10/OCD.ttl	
LinkedMDB	http://www.cs.toronto.edu/$$oktie/linkedmdb/	http://www.linkedmdb.org/snorql/
Linked Open Vocabularies (LOV)	http://lov.okfn.org/dataset/lov/	http://lov.okfn.org/dataset/lov/sparql

The datasets have various amounts and sizes. A dataset may consist of one to dozens of RDF files, while the size of RDF files also vary from several kilobytes to reach tens of gigabytes. While running identification queries [25], we split up the large datasets into small partitions. The results of identification dataset can be found in Table 3.

Table 3. Datasets statistical information

Dataset	Triple	Class	Property	Subject	Entity	Object
British National Bibliography	144.754.861	29	63	19.100.626	12.682.101	21.284.676
Citeseer	6.282.235	4	19	859.175	859.175	859.468
DBLP	116.815.316	14	27	5.593.196	5.593.196	27.045.002
DBpedia	845.852.363	2.424	63.100	50.483.918	22.669.052	105.020.130
Democratic City	1.379.344	21	57	216.147	216.127	59.875
GeoNames	169.462.379	1	26	11.341.387	11.341.387	35.267.801
Lexvo	726.674	5	19	128.945	107.517	240.334
Linked Open	65.829	9	45	10.998	8.227	9.566

Vocabularies						
LinkedGeoData	1.292.933.812	1.136	32.491	284.098.269	267.737.503	268.402.435
LinkedMDB	6.147.996	53	222	694.400	665.441	1.049.261
OCD : Cidade Democratica Ontology	822	5	12	220	177	220
Quran Ontology	1.290.773	45	82	111.004	110.946	32.454
Semantic Quran	15.741.614	10	65	1.463.769	1.446.873	497.586
Wikidata	2.269.619.296	443	14.596	262.057.530	259.212.513	182.702.002
WordNet	5.557.709	5	64	647.215	645.761	1.193.996
Yago	1.090.372.899	484.085	88.736	331.212.386	15.368.508	17.412.052

2.2. Set of Queries

At this stage, a set of queries generated based on the most classes and properties that are frequently occurred in all datasets. However, two journals ("Luzzu - A Framework for Linked Data Quality Assessment" and a journal entitled "Semantic Hadith Leveraging Linked Data Opportunities for Islamic Knowledge") whose datasets have only specific classes and properties and most are not found in other journals are excluded in evaluation since the generated queries cannot be run over those two datasets.

There are two types of queries to be created, namely star queries and chain queries. Star queries are used to get one subject that has multiple predicates and objects, while the query chain is used to retrieved some subjects and objects, where an object can be a subject to obtain other objects [11]. Each query is added a limit keyword for reducing the possibility of too long query processing time.

Star Query 1 contains four rows which consist of one class and four properties. Star Query 1 retrieves person URI, person name (label), person description (comment) and any related information (owl:sameAs).

```
Select * {
 ?s   <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://xmlns.com/foaf/0.1/Person>.
 ?s        <http://www.w3.org/2000/01/rdf-schema#label>
?label.
 ?s <http://www.w3.org/2002/07/owl#sameAs> ?sameas.
 ?s      <http://www.w3.org/2000/01/rdf-schema#comment>
?comment.
 }
 LIMIT 5
```

Star Query 2 consists of one class and three properties in three triples. Star Query 2 is used to obtain organization URI, organization name, and organization homepage.

```
select * {
 ?s   <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://xmlns.com/foaf/0.1/Organization>.
 ?s <http://xmlns.com/foaf/0.1/name> ?name.
 ?s <http://xmlns.com/foaf/0.1/homepage> ?homepage.
 }
 LIMIT 5
```

Chain Query contains two rows which consist of two properties. Chain Query retrieves subject URI, other equivalent URI, and other URI related to subject.

```
select * {
 ?s <http://www.w3.org/2002/07/owl#sameAs> ?sameas.
 ?sameas                <http://www.w3.org/2000/01/rdf-
schema#seeAlso> ?seealso.
 }
 LIMIT 5
```

5. RESULT AND DISCUSSION

The average number of triples and properties in paper 7 [20] is the highest, while the average number of triples in paper 6 is the lowest. The ratio of average number of classes and average number of triples in paper 6 is low, it implies that the amount of class diversity is quite high. Moreover, the average subjects and entities is likely the same. It can be concluded that most of triples contains rdf:type.

40% of datasets have a *SF* value less than 0.35. It means that the distribution of classes and properties are not distributed evenly. None of datatasets reach *SF* value more than 0.6, since many classes do not belongs to all datasets.

Paper 4 consisting two datasets has a high *SFQ* values. The two datasets in that journal has high number of classes and properties that are included in the set of queries such as rdf:label and foaf:name. Although average number of classes and properties in paper 5 and 7, the *SF* and *SFQ* value are low. It indicates that some classes and properties are only populated in a certain dataset.

Although Paper 10 has a high *SF* value, the *SFQ* value is the lowest amongst other journal papers. It seems that the classes and properties used in queries spread across over the dataset such as owl is sameAs and rdfs is a label as seen in Table 4.

Table 4. The average number of triples and properties

Paper	Average						SF	SFQ
	Triple	Class	Property	Subject	Entity	Object		
1	423.289.518,5	1.214,5	31.559,5	25.306.431,5	11.388.284,5	52.630.232	0,50002535	2,166666667
2	41.009.713,67	25,33333333	98	2.099.531,333	2.088.954,667	9.367.943	0,346470172	2
3	1.069.393.088	1.780	47.795,5	167.291.093,5	**145.203.277,5**	**186.711.282,5**	0,500025108	1,833333333
4	495.303.612	1.226,5	31.581,5	34.792.272	17.675.576,5	63.152.403	0,500226496	2,5
5	647.260.990,3	162.171,3333	50.633,33333	127.447.839,7	12.894.440,33	41.208.726	**0,333347652**	1,777777778
6	**690.083**	13	34,5	**108.183,5**	**108.152**	**30.047,5**	0,505865103	-
7	**1.401.948.186**	**162.317,3333**	**55.477,33333**	**214.584.611,3**	99.083.357,67	101.711.394,7	**0,333349631**	1,777777778
8	8.516.193,5	27,5	73,5	787.386,5	778.909,5	265.020	0,527769074	-
9	769.416.184,7	1.187	31.872,33333	115.307.858	100.582.647,3	136.230.122	**0,333364014**	1,333333333
10	61.548.775,5	9	23	3.226.185,5	3.226.185,5	13.952.235	0,537383178	**1,166666667**

6. CONCLUSION

16 datasets that are used in 10 Linked Data papers have been investigated in this paper. The dataset characteristics that are investigated are the number of classes, properties, entities, objects, and subjects. Moreover, the distribution of classes and properties are also considered as one of characteristics of the dataset. The main characteristics of these datasets are rdf is a type, owl is sameAs, and rdfs is a label occur in all datasets. The distribution of properties and classes is not evenly. In the future, we need to evaluate the performance of a federated engine over those datasets by using the generated queries.

7. REFERENCES

[1] Gandon, F., & Schreiber. G. (2014). RDF 1.1 XML Syntax. W3C Recom- mendation.

[2] Harris, S., & Seaborne, A. (2013). SPARQL 1.1 query language. W3C, Working Draft.

[3] Go¨rlitz, O., Thimm, M., & Staab, S. (2012). Splodge: Systematic generation of sparql benchmark queries for linked open data. *International Semantic Web Conference,* (1), 116–132.

[4] Schwarte, A., Haase, P., Hose, K., Schenkel, R., & Schmidt, M. (2011). Fedx: a federation layer for distributed query processing on linked open data," in *Extended Semantic Web Conference*. Springer, 481–486.

[5] Acosta, M., Vidal, M.E., Lampo, T., Castillo, J., & Ruckhaus, E. (2011). Anapsid: an adaptive query processing engine for sparql endpoints. *The Semantic Web–ISWC 2011*, pp. 18–34.

[6] Lynden, S., Kojima, I., Matono, A., & Tanimura, Y. (2011). Aderis: An adaptive query processor for joining federated sparql endpoints. *On the Move to Meaningful Internet Systems: OTM 2011*, pp. 808–817.
[7] Rakhmawati, N. A., Karnstedt, M., Hausenblas, M., & Decker, S. (2014). On metrics for measuring fragmentation of federation over sparql endpoints. *WEBIST,* (1), 119–126.
[8] Montoya, G., Vidal, M.E., Corcho, O., Ruckhaus, E., & Buil-Aranda, C. (2012). Benchmarking federated sparql query engines: Are existing testbeds enough. *International Semantic Web Conference*. Springer.
[9] Schmidt, M., Go¨rlitz, O., Haase, P., Ladwig, G., Schwarte, A., & Tran, T. (2011). Fedbench: A benchmark suite for federated semantic data query processing. *The Semantic Web–ISWC 2011*, 585–600.
[10] Wu, H., Fujiwara, T., Yamamoto, Y., Bolleman, J., & Yamaguchi, A. (2014). Biobenchmark toyama 2012: an evaluation of the performance of triple stores on biological data. *Journal of biomedical semantics*, 5(1), 32.
[11] Rakhmawati, N.A., Saleem, M., Lalithsena, S., & Decker, S. (2014). Qfed: Query set for federated sparql query benchmark. *Proceedings of the 16th International Conference on Information Integration and Web- based Applications & Services*.
[12] Saleem, M., Mehmood, Q., & Ngomo, A.C.N. (2015). Feasible: A feature-based sparql benchmark generation framework. *International Semantic Web Conference*. Springer.
[13] Duan, S., Kementsietsidis, A., Srinivas, K., & Udrea, O. (2011). Apples and oranges: a comparison of rdf benchmarks and real rdf datasets. Proceedings of the 2011 *ACM SIGMOD International Conference on Management of data. ACM*.
[14] Westphal, P., Stadler, C., & Pool, J. (2015). Countering language attrition with panlex and the web of data. *Semantic Web*, 6(4), 347–353.
[15] Wu, H., Villazon-Terrazas, B., Pan, J.Z., & Gomez-Perez, J.M. (2014). How redundant is it-an empirical analysis on linked datasets. *Proceedings of the 5th International Conference on Consuming Linked Data-Volume 1264*. CEUR-WS.
[16] Endris, K.M., Faisal, S., Orlandi, F., Auer, S., & Scerri, S. (2015). Irap-an interest-based rdf update propagation framework. International Semantic Web Conference (Posters & Demos).
[17] Adamou, A., d'Aquin, M., Barlow, H., & Brown, S. (2014). Led: curated and crowdsourced linked data on music listening experiences. *Proceedings of the 2014 International Conference on Posters & Demonstrations Track-Volume 1272*. CEUR-WS.
[18] De-Melo, G. (2015). Lexvo. org: Language-related information for the linguis-tic linked data cloud. *Semantic Web*, 6(4), 393–400.
[19] Debattista, J., Auer, S., & Lange, C. (2016). Luzzu-a framework for linked data quality assessment. *Semantic Computing (ICSC),* 2016 IEEE Tenth International Conference on. IEEE, pp. 124–131.
[20] Baier, J., Daroch, D., Reutter, J.L., & Vrgoc, D. (2016). Property paths over linked data: Can it be done and how to start. *COLD@ISWC*.

[21] Basharat, A., Abro, B., Arpinar, I.B., & Rasheed, K. (2016). Semantic hadith: Leveraging linked data opportunities for islamic knowledge. *LDOW@WWW*.
[22] Kurte, K.R., Durbha, S.S., King, R.L., Younan, N.H., & Vatsavai, R. (2017). Semantics-enabled framework for spatial image information mining of linked earth observation data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(1), 29–44.
[23] Hu, Y., Janowicz, K., Hitzler, P., & Sengupta, K. (2015). The semantic web journal as linked data. I*nternational Semantic Web Conference (Posters & Demos)*.
[24] Convert Geo Names RDF dump format into triples.
[25] Cyganiak, R. An RDF schema and associated documentation for expressing metadata about RDF datasets.