A generalization SPARQL federated query: An initial step towards machine-readable web of data for halal food products

A Hernawan1*, AL Sunarwidhi2, ES Prasedya3, S Widyastuti4

1Department of Informatics Technology, Faculty of Engineering, University of Mataram, Indonesia
2Department of Pharmacy, Faculty of Medicine, University of Mataram, Indonesia
3Bioscience and Biotechnology Research Centre, Faculty of Mathematics and Natural Sciences, University of Mataram, Indonesia
4Faculty of Food Technology and Agroindustry, University of Mataram, Indonesia

*Corresponding author: arihernawan@unram.ac.id

Abstract. The evolution of webs from web of documents to a web of data allows machines to read and connect structured data on the Web, known as Linked Data. This enables machine learning to use linked data in ways that were previously thought to be impossible to do. Machine learning and linked data research are relatively thoughtless. As a result, the primary goal of this research is to present ARQL, an extension of SPARQL 1.1 and to cluster Linked Data derived from online web sources. We provide state of the art ARQL features to allow machine to directly query from online RDF, JSON-LD or JSON API. Here we describe how ARQL web service is used to query Web of Data development and generalization of SPARQL federated query, two initial steps for the development of halal food mobile application tied up with ARQL.

Keywords: Linked Data; SPARQL; Halal Food Product

1. Introduction
In recent years, the web has evolved from web of document to web of data, which was Tim Berners-Lee primary goal in designing the web [1]. The Web evolves from a global information space of linked documents, to one where both documents and data are linked, providing a solid foundation for this evolution to be best practices for publishing and connecting structured data on the web known as Linked Data. The main difference between HTML Web page and Linked Data is that HTML follows links between HTML pages, meanwhile Linked Data allows user to navigate between any data source by following RDF links. This enables user to begin with a single data source and then navigate through a potentially infinite Web of data sources linked by RDF links. While more and more structured data is being published on the web in accordance with the Linked Data principles [2], an important question that needs to be addressed is how to efficiently access and query this expanding body of knowledge. The most well-known query language for Linked Data is SPARQL. Unfortunately, the majority of SPARQL implementation requires the data to be prepared ahead of time, either in main memory or repository as SPARQL endpoint. Nonetheless, Linked Data exists in the Web in various forms, even HTML Web page can contain RDF data through RDFa or RDF data, which may be dynamically created by Web Services.
The Worldwide Web is a global information space where Open Data is collected, freely used and easily distributed. When combined with open data, Linked Data can be referred to as Linked Open Data (LOD). The ease of use to this data could lead into numerous research and applications. Machine Learning (ML) techniques will be required for mining the Web of Data in knowledge-intensive applications that use LOD [3]. As a result, unsupervised machine learning techniques for clustering RDF document becomes a mandatory requirement.

In this research, Halal food products were chosen due to the large impact these products have on health and well-being. Halal is a Shariah-compliant code of conduct that applies to every activity performed by a Muslim [4]. It is becoming more difficult to identify halal components in processed food due to the advance in science and technology of food processing. To determine halal status of the compounds in the processed food, we need to trace back to the original source. Moreover, to design culturally acceptable foods and optimize healthy food choices, an integrated transdisciplinary approach is required [5].

Institute for Foods, Drugs and Cosmetics Indonesian Council of Ulama, known as LPPOM MUI, is an authorized institution in Indonesia, which has the duty to supervise halal foods distribution in Indonesia. The institution provides a website where users can search for a halal certificate of product based on either product name or company name [6]. The halal certificate includes the registration number, product name, company name and expiration date. The user can also download the entire list of documents in PDF format. Unfortunately, data format that has been provided is not Linked Data, making it difficult for machine to read and combine with other data sources. Therefore, a new system is needed to address this issue. In order to exploit Web of Data from multidisciplinary areas to provide halal information of food product, an initial step is needed. Here we describe how ARQL web service to query Web of Data was developed and how generalization ARQL federated query was performed, which allows user to transform and query JSON API as same as Linked Data.

2. Materials and Methods
The data sources used in this study were LPPOM MUI Database, Open food fact, DBpedia, and European Commission Food Additives Database. These data sources contain valuable information that is stored in various formats including RDF, RDFa and JSON. This research generalizes SPARQL federated query to directly exploit and combine the data. The proposed federated query does not require the data to be stored locally. Direct transformation from JSON to Graph was developed as an extend to JSON-LD.

3. Results and Discussion
The development of machine-readable web of data for halal food products will be useful to facilitate people to identify the halal status of compounds in food products. The initial step of the development
of the app utilized several data sources, including LPPOM MUI Database, Open food fact, DBpedia, and European Commission Food Additives Database. Institute for Foods, Drugs and Cosmetics Indonesian Council of Ulama (LPPOM MUI) is an authorized institution to supervise halal food product that is distributed in Indonesia [7]. Currently LPPOM MUI data source has provided halal product information services via SMS gateway, web and mobile application. They provide data related to product name, halal certificate number, producer name and date expiry. These data was then exploited in a form of Web page, PDF file or JSON API.

The second data source used, the Open food fact, is a free and open database of food products from around the world. Open food facts provide information about food products including food description, manufacture ingredients and nutrition. It presents a halal label for some food product but the clarity of the halal status is still questionable. They provide service where users could use their open data by downloading bulk data in CSV and RDF format. Another option is JSON API service, which could facilitate the search of food detail information. The RDF bulk data contains less information compared to JSON and HTML Webpage. Open food fact graph contains three classes: Food Product, Ingredient and Food Information.

The third data source was DBpedia, which is a community where structured information from Wikipedia could be obtained. DBpedia provides a SPARQL endpoint, which allows other users to search by using queries on Wikipedia and also connect DBpedia dataset to other datasets. Lastly, the European Commission Food Additives Database was also used as a data source in this study. This database serves as a tool to inform users about the food additives that has been approved for use in the EU area [8]. European Economic Community (EEC) initiated E-Number code to represent list of food additives. E-Number is commonly used by the food industry in manufacture of various food products and up to now universally adopted by the food industry around the world.

Several related studies in the literature related to Querying Linked Data, Machine Learning and food domain has been conducted. For instance, Fafalios et al., [9] introduced SPARQL-LD that can query a dataset coming from the partial results of query or RDF data from Web Services. SPARQL-LD extends the applicability of the service operator and enables the query of any HTTP Web source containing RDF data. This extension does not require the named graphs to be declared, thus a user can even obtain a dataset returned by a portion of the query when URI is derived at query execution time. Moreover Rakhmawati et al., [6] proposed a halal food vocabulary, transforming related halal food data, integrating Linked Data and building a web application to predict halal status of an uncertified halal product. Proposed halal vocabulary contains a set of halal information including certification code, expired date of certification, and halal organization. Dataset was gathered from LPPOM MUI, E-Number, PubChem and Mesh.

Predicting the halal status of food has been done in several other studies [10], [11], where the examined E-number for assisting the Muslim community in verifying the halal status of substances used as the ingredients for the manufacture of a food and beverage product. They built an ontology and database based on PubChem and Chebi dataset. Kasim et al., [10] provided places where a user can check a halal of the status of E-Code. Meanwhile Hashim., [11] designed the halal framework, for tracing a halal status of flavoring ingredient which aims to increase a halal certification process. Another source known as Apache Jena is a free and open-source Java framework for building Semantic Web and Linked Data applications [12]. It has SPARQL 1.1 compliant engine known as ARQ. ARQ support remote federated queries and free text search.

Traditionally, most SPARQL endpoint and web services need to have data available in advance. Data are stored locally in repository or database following by specific ontology or schema. Our proposed ARQL will act as SPARQL endpoint that querying Web of Data and transform JSON in such a way that will work with Linked Data. ARQL implement all new features from its core and pairing with other application is not necessary.

ARQL is originally from Apache Jena. Prior work on SPARQL-LD has been done in version 2.13.0 and major updates has been done since then. Most code package and structure in recent version are completely different. More work was added to implement proposed SPARQL 1.1 Federated Query.
ARQL Federated Query extend service operator functionality. Default service operator functionality are kept unchanged. Modification is solely down to the evaluation of a graph from JSON or RDF graph served by a web resource. The main program will build RDF graph from every external data source received from URL service. Since JSON is not format suitable to build RDF graph, new block of code is created to transform JSON into graph. ARQL allows user to directly exploit RDF, RDFa and JSON from web. Here we show example queries that demonstrate the functionality offered by ARQL:

3.1. Querying RDFa
Web page BBC food recipes contain many varieties recipes options. It categorized based on calories consumption, culture and other dietary. Underneath food recipes HTML page contains RDFa document that allows machine to read and possibility processed with other data source. ARQL can query RDFa document directly from Food recipes Web pages.

```sql
SELECT * 
WHERE { 
  SERVICE <https://www.bbc.co.uk/food/recipes/croque_madame_76205> { 
    ?s ?p ?o 
  } 
}
```

**Figure 2.** RDFa query on BBC web pages

3.2. Querying JSON
ARQL transforms JSON hierarchal level into graph RDF model. Complex linked data that modeled as an RDF graph could have loop connection inside graph. Compare to complex RDF document, JSON structure is translated into a tree in graph model. Since all data stored in same RDF model, ARQL can query linked data without issues. The example of this query is shown in Figure 3.

```sql
SELECT ?foodCode ?additives 
WHERE { 
  BIND(STR("089686041705") as ?foodCode). 
  BIND(REPLACE(str("https://world.openfoodfacts.org/api/v0/product/__foodCode__.json"), 
    "__foodCode__", str(?foodCode), "i") as ?x) BIND(URI(?x) as ?service). 
  SERVICE ?service { 
    ?o_s json:additives_original_tags[ ?p_additives ?additives]. 
  } 
}
```

**Figure 3.** JSON query on Open food fact API endpoint

4. Conclusion
We have presented ARQL that has a new query set to query Web of data using generalization SPARQL 1.1 federated query. The proposed method shows interconnected SPARQL endpoint, RDF dump and JSON document that is available on Web. The future semantic web will also contribute to create massive global open database for linked data. Knowledge intensive application will have an option to exploit directly from Web instead of duplicate data to their local database. This approach will save computing resource to store RDF data locally. Moreover, future development of ARQL will open possibility to research and studies of machine learning together with semantic web.

Acknowledgement
We would like to thank all parties that had contributed to this research.
References

[1] T. Berners-Lee, “Linked Data - Design Issues,” 2006. [Online]. Available: https://www.w3.org/DesignIssues/LinkedData.html.

[2] C. Bizer and F. U. Berlin, “Linked Data : The Story so Far Linked Data - The Story So Far,” no. July, 2009, doi: 10.4018/jswis.2009081901.

[3] S. Eddamiri, E. M. Zemmouri, and A. Benghabrit, “An improved RDF data Clustering Algorithm,” vol. 00, pp. 1–10, 2019, doi: 10.1016/j.procs.2019.01.038.

[4] G. Rezai, Z. Mohamed, and M. N. Shamsudin, “Assessment of Consumers ’ Confidence on Halal Labelled Manufactured Food in Malaysia,” vol. 20, no. 1, pp. 33–42, 2012.

[5] M. B. Cole, M. A. Augustin, M. J. Robertson, and J. M. Manners, “The science of food security,” npj Sci. Food, no. July 2017, pp. 1–8, 2018, doi: 10.1038/s41538-018-0021-9.

[6] N. A. Rakhmawati, J. Fatawi, A. C. Najib, and A. A. Firmansyah, “Linked open data for halal food products,” J. King Saud Univ. - Comput. Inf. Sci., no. xxxx, 2019, doi: 10.1016/j.jksuci.2019.04.004.

[7] G. Atmadi and S. R. W. Widati, “Strategi Pemilihan Media Komunikasi LPPOM MUI Dalam Sosialisasi & Promosi Produk Halal di Indonesia,” J. Al Azhar Indonesia. Seri Pranata, no. 2, pp. 87–97, 2013.

[8] EU, “European Commission Food Additives Database.” [Online]. Available: https://ec.europa.eu/food/safety/food-improvement-agents/additives/database.

[9] P. Fafalios, T. Yannakis, and Y. Tzitzikas, “Querying the Web of Data with SPARQL-LD,” pp. 1–12, 2016.

[10] S. Kasim, N. A. Omar, M. F. M. Fudzee, A. A. Ramli, M. A. Salamat, and H. Mahdin, “The Development of Ontology from Multiple Databases The Development of Ontology from Multiple Databases,” Int. Res. Innov. Summit, 2017, doi: 10.1088/1757-899X/226/1/012108.

[11] S. F. M. Hashim, J. Salim, S. A. Noah, and W. A. W. Mustapha, “A Framework for Tracing the Flavouring Information to Accelerate Halal Certification,” vol. 9, no. 2, 2017.

[12] The Apache Software Foundation, “Apache Jena.” [Online]. Available: https://jena.apache.org/.