Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 197 (2022) 362–369

Procedia Computer Science

www.elsevier.com/locate/procedia

Sixth Information Systems International Conference (ISICO 2021)

Indonesia COVID-19 cases report using Linked Open Data

Nur Aini Rahkmawati*, Adam Akbar, Bramantyo Adhilaksono, Fikri Baharuddin, Rahmat Hidayat

Departement of Information Systems, Institut Teknologi Sepuluh Nopember, Jl. Raya ITS, Keputih, Sukolilo, Surabaya, 60111, Indonesia.

Abstract

Coronavirus disease is a worldwide pandemic. The need for accurate data and information become an important thing in this pandemic situation. In Indonesia, the government provides an official website for displaying COVID-19 spread statistics. However, the data provided does not follow the 5-star open data. As a result, the data is not reusable and integrated easily into another dataset and application. In this paper, we proposed an RDF vocabulary for presenting COVID-19 data in Indonesia. In addition, two queries are presented as an example for using our vocabulary and dataset as part of Linked Open data movement.

Keywords: : Linked Open Data; COVID-19; RDF; SPARQL

1. Introduction

In this worldwide pandemic of coronavirus disease (COVID-19) crisis, accurate data information is important. Since it was first discovered in Wuhan (Hubei Province, China), COVID-19 has infected hundreds of millions of people worldwide [1]. Covid is caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [2] which continues to mutate and give rise to new variants [3]. Most countries have implemented responsive measures as a reaction to COVID-19 [4]. Governments around the world also publish COVID-19 related data daily. These published COVID-19 data may unveil hidden patterns and aid in developing a better understanding of the pandemic [5].

* Corresponding author.
E-mail address: nur.aini@is.its.ac.id

10.1016/j.procs.2021.12.151

Most countries are motivated to collect data and provide daily reports on the virus spread, the number of infected, recovered, and deceased persons for connection to other country data [6]. On 15 May 2021, 1,731,652 Indonesian people has infected COVID-19 and 47,716 people deaths [7]. The Indonesian government provides a website https://covid19.go.id/ [8], which displays the map of the distribution of the COVID-19 in Indonesia, such as the number of positive confirmed cases in a day, case in a province, symptoms, etc. Different formats for data management also leads to a problematic situation for exchanging information for COVID-19 spreads care [9]. However, it does not follow the 5-star open data [10]. The number of data is increasing daily on the website, so it needs to create a good data representation. Furthermore, data integration is also needed to facilitate the development of existing data.

Nagai et.al [11] presented the frequency of COVID-19 related words from several websites in Japan using Linked Open Data, while we present Linked Open Data for the Indonesian COVID-19 case. Similar to our work, Ulahannan et.al [12] proposed a visualization for Kerala, India case, however the dataset does not follow Linked Open Data principle. Linked Data employs the Resource Description Framework (RDF) and the Hypertext Transfer Protocol (HTTP) to publish structured data on the Web and to connect data between different data sources [13]. Linked Open Data is also Linked Data, released under an open license, which does not impede its reuse for free [14]. RDF is an XML syntax for expressing metadata and schemas in a form that is both humanly readable and machine-readable [15]. Meanwhile, a turtle document is a textual representation of an RDF graph [16].

We design a vocabulary that accommodates the number of cases by province in Indonesia and symptoms information of COVID-19. Moreover, collected data from https://covid19.go.id/ are transformed into RDF, and the RDF data are connected to DBpedia. Also, a set of queries are performed over the data as a showcase.

2. Methodology

We initially collect raw data from the source. Based on the collected data, we design a vocabulary that extends from a popular vocabulary. Our methodology can be explained as follows:

2.1. Dataset

In this paper, we focused on design two-part of data in covid19.go.id [8] namely: the number of cases by province and the percentages of COVID symptoms. We collect raw data on day 4 and 5, May 2021, in JSON extension. Raw data can be accessed at [17]. the number of cases by province data consists of *datePosted*, *province name*, *totalConfirmedCase*, *totalActiveCase*, *totalPatientHealed*, and *totalPatientDeath*. Moreover, the percentages of COVID symptoms data contains *datePosted*, *symptomName*, *totalPatientWithSymptomPositivePercentage*. A summary of the data can be seen in Table 1. A summary of the data can be seen in Table 1. The sample of the number of cases by province dataset is presented in Table 2, while the number of COVID symptoms can be seen in Table 3.

Table 1. Summary of data.

Data	Property
cases by province	*datePosted*
	provinceName
	totalConfirmedCase
	totalActiveCase
	totalPatientHealed
	totalPatientDeath
Percentage of symptoms	3

Table 2. Dataset of the number of COVID-19 cases by province in Indonesia.

#	Date posted	Province name	Total confirmed case	Total active case	Total patient healed	Total patient death
1	2021-05-04	Jakarta	410400	6657	397079	6704
2	2021-05-05	Jakarta	411573	6527	398317	6729
3	2021-05-04	West Java	282631	30597	248276	3758
...	...	...	...	...	...	...
18	2021-05-05	Riau	46061	5185	39735	1141
19	2021-05-04	Special Region of Yogyakarta	39824	3814	35045	965
20	2021-05-05	Special Region of Yogyakarta	40140	3478	35681	981

Table 3. Dataset of the percentages of sympthons in positive patients.

#	datePosted	symptomName	PercentageofSymptomPositivePatient
1	2021-05-04	Cough	63.2
2	2021-05-05	Cough	63.2
3	2021-05-04	Fever	35.9
...	...	...	...
18	2021-05-05	Colic	5.7
19	2021-05-04	Diarrhea	5.5
20	2021-05-05	Diarrhea	5.5

2.2. Vocabulary design

Based on the dataset described earlier, we develop two classes: (i) *Covid_Case_In_Indonesia_Province* for cases by province data, and (ii) *Covid19_Symptom_Statistics* for the percentage of symptom data. We use *SpecialAnnouncement* class [16] from Schema.org as the base class for *Covid_Case_In_Indonesia_Province* and Covid19_Symptom_Statistics class. We extend that class by adding few properties to accommodate all properties on the dataset. SpecialAnnouncement itself is motivated by the COVID-19 pandemic, and it is aligned well with our goal.So, properties for *Covid_Case_In_Indonesia_Province* class are: (i) *diseaseSpreadStatistics*, (ii) *datePosted*, (iii) *spatial*, (iv) *totalConfirmedCase*, (v) *totalActiveCase*, (vi) *totalPatientHealed*, (vii) *totalPatientDeath*, and (viii) *disease*. Properties for *Covid_Case_In_Indonesia_Province* class are: (i) *diseaseSpreadStatistics*, (ii) *datePosted*, (iii) *spatial*, (iv) *symptom*, (v) *disease*, and (vi) *totalPatientWithSymptomPositivePercentage*. To link our data with another open data, *spatial*, *symptom*, and *disease* properties from two classes are linked to DBpedia data. All proposed classes are represented in Fig. 1 as class map.

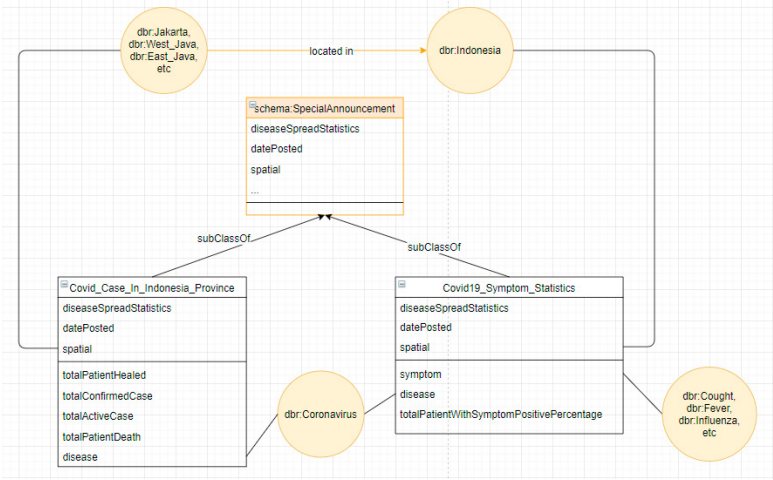

Fig. 1. Indonesia COVID-19 Linked Open Data class map.

3. Turtle example and visualization

3.1. Vocabulary implementation

Based on the design from the previous section, we then implement the vocabulary in Turtle format. We generate the graph representation of the implemented vocabularies. The generated RDF graphs for the number of cases by province and percentage by symptom are presented in Fig. 2 and Fig. 3, respectively.

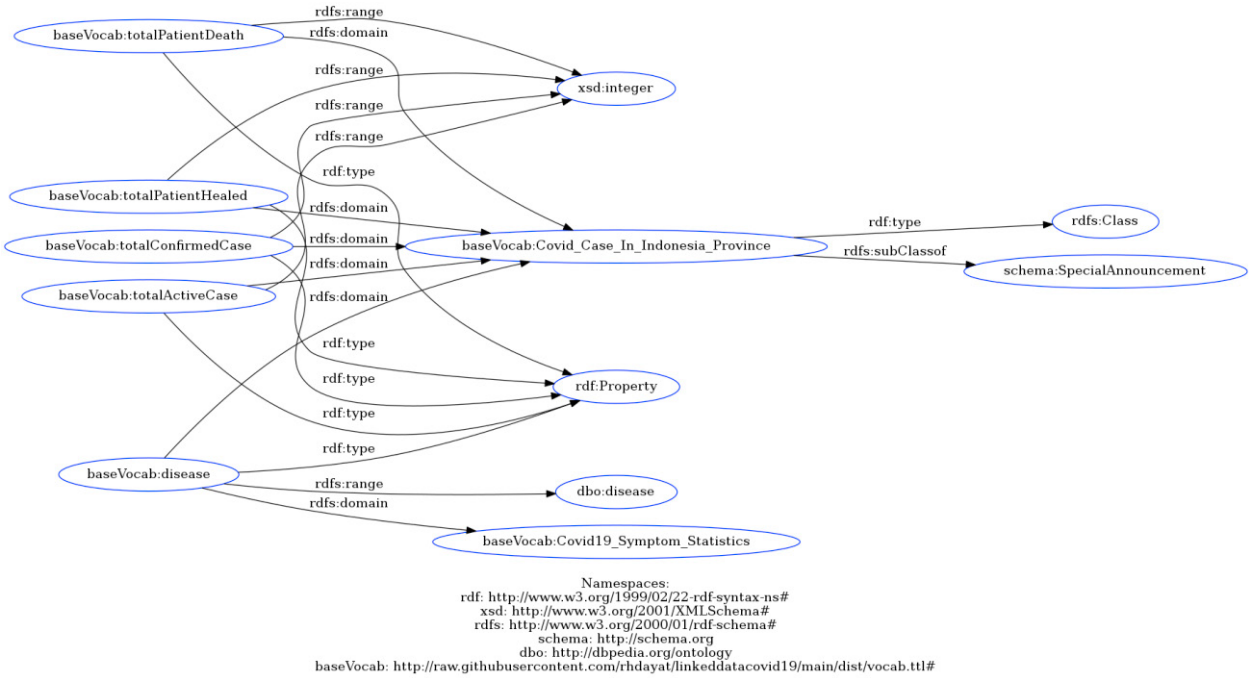

Fig. 2. RDF graph for *Covid_Case_In_Indonesia_Province* class.

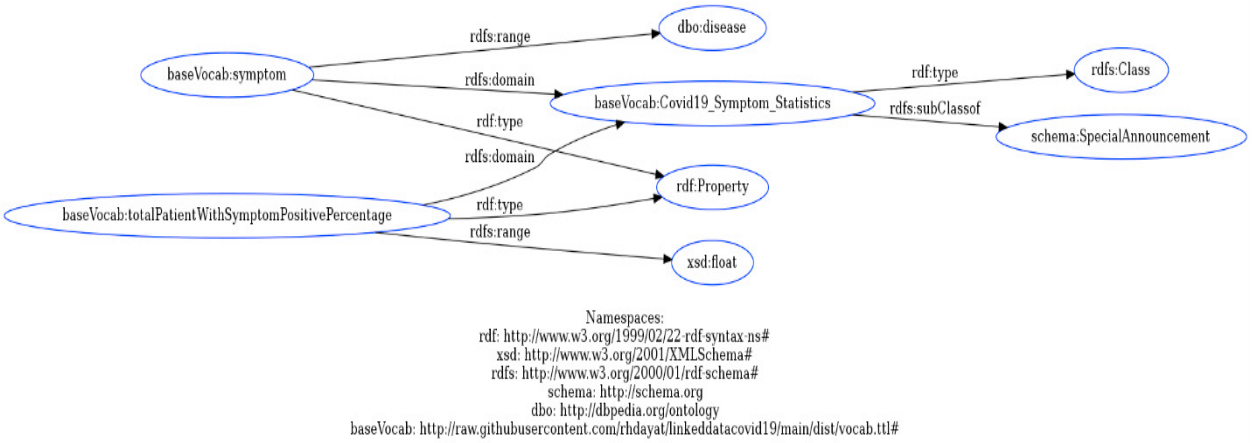

Fig. 3. RDF graph for *Covid19_Symptom_Statistics* class.

3.2. Data conversion

The next step is to convert the data used into Turtle format. As explained in the previous section, we only converted 10 data from each dataset since the website only exposed data for the current date and DKI Jakarta is the capital of Indonesia. As a sample, an RDF graph is generated from the converted data. RDF graph for the number of cases in DKI Jakarta as of May 4th, 2021, is shown in Fig. 4. The second dataset used is the percentage by symptom. As a sample, we only show cough data retrieved on May 4th, 2021. The converted data and RDF graph for cough symptom is presented in Fig. 5.

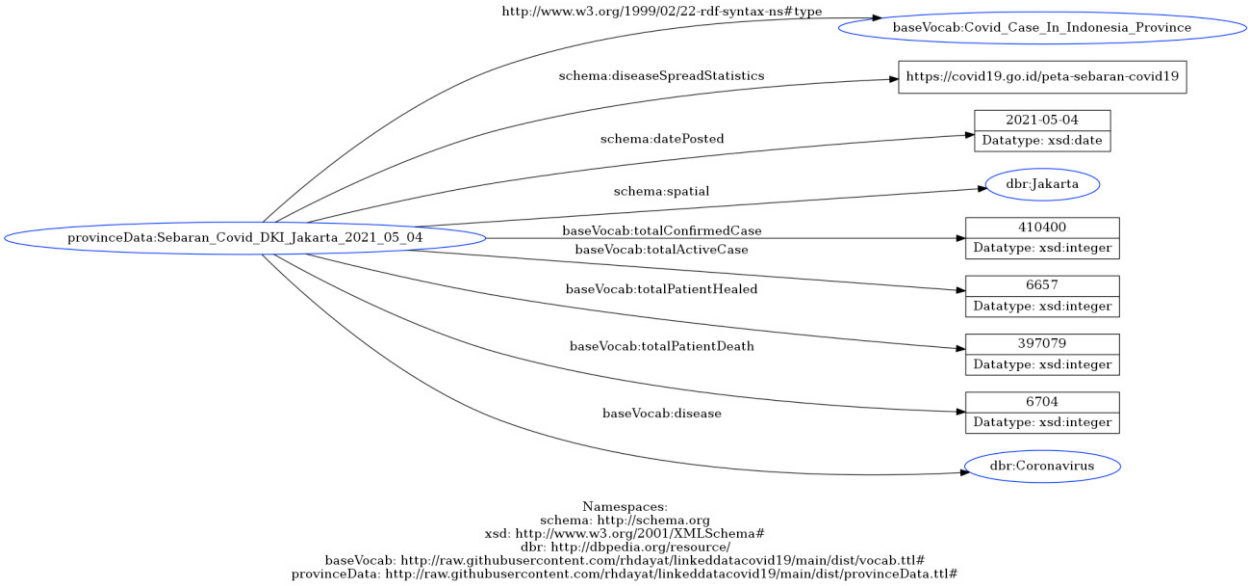

Fig. 4. RDF graph of sample number cases by province data.

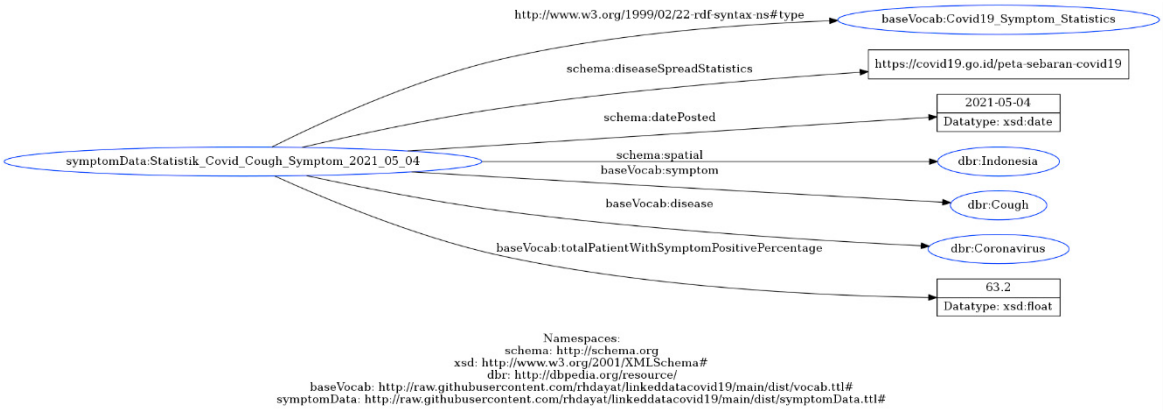

Fig. 5. RDF graph of sample percentage by symptom data.

4. Querying the Linked Open Data

We translate all data from table-based to RDF Turtle, based on the designed class described in the previous section. Covid_Case_In_Indonesia_Province class for cases by province, and Covid19_Symptom_Statistics class for percentage by symptom. To query linked data, we use SPARQL as query language. For example, we want to know the total of active cases each day based on the dataset. The totalActiveCase is calculated from Covid_Case_In_Indonesia_Province based on their datePosted. The totalActiveCase is the sum of Covid_Case_In_Indonesia_Province based on their datePosted multiplied by totalPatientWithSymptomPositivePercentage from Covid19_Symptom_Statistics that has the same datePosted value. Example of queries for querying those data are shown in Fig. 6 and Fig. 7. The query results are presented in Table 4 and Table 5. The complete query result can be accessed at [17].

```
PREFIX dbr: <http://dbpedia.org/resource/>
PREFIX schema: <http://schema.org/>
PREFIX baseVocab: <http://raw.githubusercontent.com/rhdayat/linkeddatacovid19/main/dist/vocab.ttl#>

SELECT ?datePosted (sum(?totalActiveCase) as ?nationalTotal)
FROM <http://raw.githubusercontent.com/rhdayat/linkeddatacovid19/main/dist/provinceData.ttl#>
{
    ?prov a baseVocab:Covid_Case_In_Indonesia_Province ;
        baseVocab:totalActiveCase ?totalActiveCase ;
        schema:datePosted ?datePosted .

}
GROUP BY ?datePosted
ORDER BY desc(?datePosted)
```

Fig. 6. Example of query total active cases each day.

```
PREFIX dbr: <http://dbpedia.org/resource/>
PREFIX schema: <http://schema.org/>
PREFIX baseVocab: <http://raw.githubusercontent.com/rhdayat/linkeddatacovid19/main/dist/vocab.ttl#>

SELECT ?symptom ?datePostedCase (?nationalTotal * ?totalPatientWithSymptomPositivePercentage / 100 as ?
SymptomTotal)
FROM <http://raw.githubusercontent.com/rhdayat/linkeddatacovid19/main/dist/symptomData.ttl#>
FROM <http://raw.githubusercontent.com/rhdayat/linkeddatacovid19/main/dist/provinceData.ttl#>
{
  ?symptom a baseVocab:Covid19_Symptom_Statistics ;
    baseVocab:totalPatientWithSymptomPositivePercentage ?totalPatientWithSymptomPositivePercentage ;
    schema:datePosted ?datePostedCase

    {
      SELECT (sum(?totalActiveCase) as ?nationalTotal)
      {
        ?prov a baseVocab:Covid_Case_In_Indonesia_Province ;
          baseVocab:totalActiveCase ?totalActiveCase ;
          schema:datePosted ?datePostedCase .
      }
      GROUP BY ?datePosted
      ORDER BY desc(?datePostedCase)
    }
}
ORDER BY ?symptom
```

Fig. 7. Example of query total active cases each symptom each day.

Table 4. Total active cases each day query result.

datePosted	nationalTotal
“2021-05-05”^^<http://www.w3.org/2001/XMLSchema#date>	63243
“2021-05-04”^^<http://www.w3.org/2001/XMLSchema#date>	64188

Table 5. Total active cases each symptom each day query result.

Symptom	datePostedCase	SymptomTotal
<http://raw.githubusercontent.com/rhdayat/linkeddatacovid19/main/dist/symptomData.ttl#Statistik_Covid_Colic_Symptom_2021_05_04>	“2021-05-04” ^^<http://www.w3.org/2001/XMLSchema#date>	“7263.567” ^^<http://www.w3.org/2001/XMLSchema#float>
<http://raw.githubusercontent.com/rhdayat/linkeddatacovid19/main/dist/symptomData.ttl#Statistik_Covid_Colic_Symptom_2021_05_05>	“2021-05-05” ^^<http://www.w3.org/2001/XMLSchema#date>	“7263.567” ^^<http://www.w3.org/2001/XMLSchema#float>
<http://raw.githubusercontent.com/rhdayat/linkeddatacovid19/main/dist/symptomData.ttl#Statistik_Covid_Cough_Symptom_2021_05_04>	"2021-05-04" ^^<http://www.w3.org/2001/XMLSchema#date>	"80536.4" ^^<http://www.w3.org/2001/XMLSchema#float>
...		

5. Conclusion

We implement the LOD principles for Indonesia Covid-19 data which are retrieved from the Indonesia government official website. The proposed vocabulary extends *SpecialAnnouncement* class from Schema.org. Two new classes are added based on the Indonesia Covid-19 data. Following the vocabulary, data are generated in Turtle format and RDF graphs. Moreover, the generated data are linked to the DBPedia dataset. We use SPARQL queries to demonstrate how to access the data that another machine can access for many purposes.

References

[1] World Health Organization (2021) "Coronavirus Disease (COVID-19) Situation Reports." [Online]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports. [Accessed 15 May 2021].

[2] World Health Organization. (2021) "Naming the coronavirus disease (COVID-19) and the virus that causes it." [Online]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it. [Accessed 15 May 2021].

[3] Cao, C., L. He, Y. Tian, Y. Qin, H. Sun, W. Ding, L. Gui, and P. Wu. (2021) "Molecular epidemiology analysis of early variants of SARS-CoV-2 reveals the potential impact of mutations P504L and Y541C (NSP13) in the clinical COVID-19 outcomes." *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases* **92**: 104831.

[4] Qundus, J. A., R. Schäfermeier, N. Karam, S. Peikert, and Adrian Paschke. (2021) "ROC: An Ontology for Country Responses towards COVID-19." *arXiv:2104.07345v1*.

[5] Santipantakis, G. M., G. A. Vouros, and C. Doulkeridis. (2021) "Coronis: Towards Integrated and Open COVID-19 Data." *Open Proceedings*.

[6] Santipantakis, G. M., G. A. Vouros, and C. Doulkeridis. (2020) "Towards Integrated and Open COVID-19 Data." *arXiv:2008.04045*.

[7] World Health Organization. (2021) "Coronavirus disease - Answers." [Online]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/coronavirus-disease-answers. [Accessed 15 May 2021].

[8] Satuan Tugas Penanganan COVID-19. (2020) "Peta Sebaran COVID-19 | Covid19.go.id [Title in English: COVID-19 Distribution Map | Covid19.go.id]." [Online]. Available: https://covid19.go.id/peta-sebaran-covid19. [Accessed 05 May 2021].

[9] Pal, Moumita, Ranjana Ray, Prasenjit Maji, and Antara Panja. (2021) "Remote Patient Monitoring during pandemic caused by COVID-19 using Semantic Web Technologies." *Journal of Physics: Conference Series* **1797(1)**.

[10] Janowicz, Krzysztof, Pascal Hitzler, Benjamin Adams, Dave Kolas, and Charles Vardeman II. (2014) "Five stars of Linked Data vocabulary use." *Semantic Web* **5**: 3.

[11] Nagai, Yuki, Tetsuya Oda, Nobuki Saito, Aoto Hirata, Masaharu Hirota, and Kengo Katayama. (2020) "Approach of a Japanese Co-Occurrence Words Collection Method for Construction of Linked Open Data for COVID-19." in *IEEE 9th Global Conference on Consumer Electronics (GCCE)*.

[12] Ulahannan, J. P., N. Narayanan, N. Thalhath, P. Prabhakaran, S. Chaliyeduth, S. P. Suresh, M. Mohammed, E. Rajeevan, S. Joseph, A. Balakrishnan, J. Uthaman, and M. Kar. (2020) "A citizen science initiative for open data and visualization of COVID-19 outbreak in Kerala, India." *Journal of the American Medical Informatics Association : JAMIA* **27 (12)**: 1913–1920.

[13] Bizer, C., T. Heath, K. Idehen, and T. Berners-Lee. (2008) "Linked Data on the Web (LDOW2008)." in *Proceedings of the 17th international conference on World Wide Web*, Beijing, China.

[14] Lee, T. Berners. (2006) "Linked Data." [Online]. Available: https://www.w3.org/DesignIssues/LinkedData.html. [Accessed 21 May 2021].

[15] Karvounarakis, Gregory, Sofia Alexaki, Vassilis Christophides, Dimitris Plexousakis, and Michel Scholl. (2002) "RQL: a declarative query language for RDF." *Proceedings of the 11th international conference on World Wide Web*.

[16] Beckett, D., T. Berners-Lee, E. Prud'hommeaux, and G. Carothers. (2014) "RDF 1.1 Turtle." *World Wide Web Consortium*.

[17] Hidayat. R., A. Akbar, F. Baharuddin, B. Adhilaksono, and N. A. Rakhmawati. (2021) "Linked Open Data COVID-19 in Indonesia | Zenodo." [Online]. Available: https://zenodo.org/record/5144334. [Accessed 29 July 2021].