2020 Asia–Pacific Statistics Week
A decade of action for the 2030 Agenda: Statistics that leaves no one and nowhere behind
15-19 JUNE 2020 | Bangkok, Thailand

Linked Open Data Implementation for Integrated Dissemination

Sulisetyo Puji Widodo [1]; Wa Ode Zuhayeni Madjida [1]; I Komang Yudi Hardiyanta [1]; Brilian Surya Budi [1]

[1] Badan Pusat Statistik

Abstract:

BPS - Statistics Indonesia disseminate official statistics on a web portal with various formats such as tables, Microsoft excel, e-books, and Application Programming Interface (APIs). The data consists of various subjects, such as social and population, economy and trade, as well as agriculture and mining. Data and information presented play an important role in evaluation and monitoring of Sustainable Development Goals (SDGs).

On other hand, today, the need for data increase continuously, not only for data availability but also for ease of data processing from various sources. However, to process data such as filtering, aggregating, integrating, including analysing and visualizing data from various sources need an effort and time consuming. This is because we have to unify different data structure and vocabulary from existing statistical data. If all information is open and uses the same format in disseminate, then we able to integrate all information from various sources.

Linked Open Data (LOD) is a technology that utilizes Open Data and allows extracting metadata in related documents or data. Then based on similarity in structure and vocabulary, LOD forms a relationship between data and metadata. If standard vocabulary is not available, we define a new vocabulary in consideration of linking with other vocabulary. Thus, the link between data can be obtained easily and make complex data processing becomes easier and more efficient.

LOD can be applied to SDGs-related data by utilizing relationships between SDGs indicators. So that all users, including stakeholders, can process data for supporting SDGs from various sources more easily. In this paper, we apply the use of LOD to statistical data specifically related to SDGs indicators.

Keywords: Open Government; Open Data; RDF; SPARQL; RDF Data Cube Vocabulary

1. Introduction:

Badan Pusat Statistik (BPS - Statistics Indonesia) is a government agency that responsible to collect and present basic statistical data in Indonesia. To meet user needs, BPS builds a web portal to disseminate the data. The data is served based on some category such as poverty, education, consumption data, etc. This also can support the achievement of Sustainable Development Goals (SDGs).

The web portal is also a BPS effort to realize Open Government. Open government provides access and licenses to public to be able to use and store data or information produced on internet by government without limitation on their use or distribution and even combine it with other data sources (Bauer. F, & Kaltenböck. M, 2012). For this reason, BPS applies Open Data concept to realize Open Government. Open data is a concept where data must be freely available for everyone to use and republish as desired, without limitation of copyright, patents, or other control mechanisms (Auer, S. R., et al., 2007). Open Data on a web portal can be seen from data availability in various forms, where user can view, copy, download as well as reuse the data as desired by user. The data format can be tables, Microsoft Excel, e-books, or API.

Nowadays, the need for data access is increasing, not just for open access to data, but also how the data can be connected and compared with other data sources based on certain metadata. Tim Berners-Lee established five level in applying open data levels. Level 1 is a condition that make our data available on the web (whatever format) under an open license. Level 2 makes it available as structured data (e.g., Excel instead of image scan of table). Level 3 prepare the data available in a non-proprietary open format (e.g., CSV instead of Excel). Using Universal Resource Identifier (URI) to denote things, so that people can point at our data is the condition at level 4. The highest level, level 5, make our data can link to other data to provide context. Figure 1 illustrates this tiers.

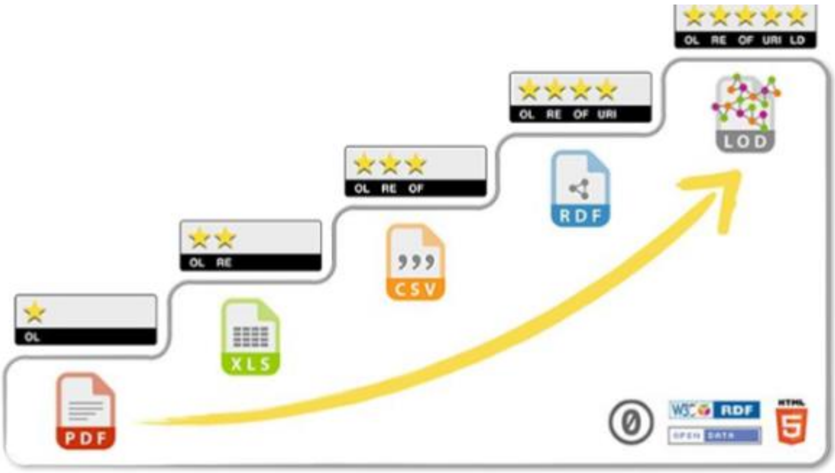


Figure 1. Open Data Level by Tim Berners-Lee

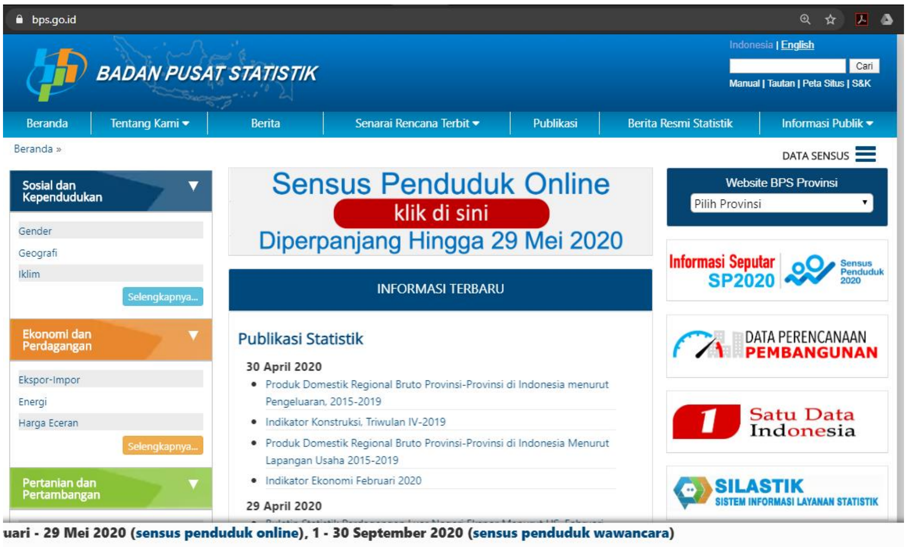
Linked Open Data (LOD) is a way to utilize Open Data and similarity of metadata so that data can be interconnected (Blaney. J, 2017). LOD uses a standard called Resource Description Framework (RDF). RDF is a recommendation of the World Wide Web Consortium (W3C) to express data into uniform structures and vocabulary. These structures allow us to analyse large volumes of data much more easily and to make comparability between data with same metadata from various source.

BPS currently has not implemented LOD in disseminating its data. This paper provides recommendations on how to implement LOD on BPS web portal. This paper also analyses how the LOD between NSOs can be interconnected so that they can realize integrated analysis.

2. Methodology:

2.1. BPS Data Portal

BPS builds a web portal as a media to disseminate data. BPS web portal can be accessed at https://www.bps.go.id/. Figure 2 shows the initial page of the BPS web portal.



Publikasi Statistik

30 April 2020

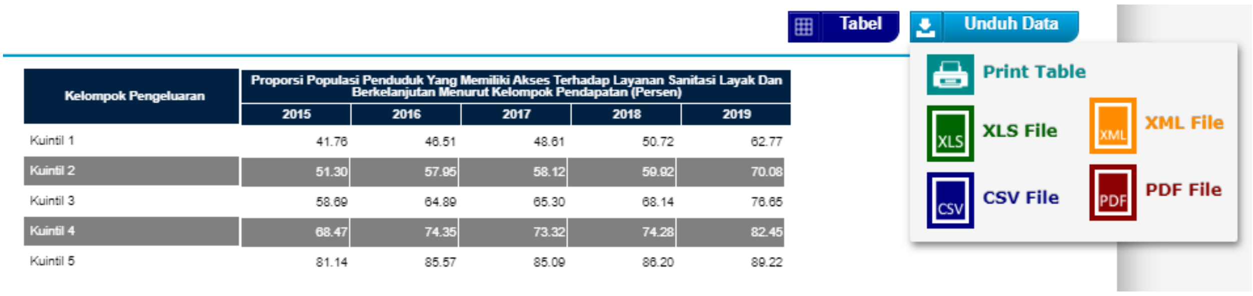
- Produk Domestik Regional Bruto Provinsi-Provinsi di Indonesia menurut Pengeluaran, 2015-2019
- Indikator Konstruksi, Triwulan IV-2019
- Produk Domestik Regional Bruto Provinsi-Provinsi di Indonesia Menurut Lapangan Usaha 2015-2019
- Indikator Ekonomi Februari 2020

29 April 2020

uari - 29 Mei 2020 (sensus penduduk online), 1 - 30 September 2020 (sensus penduduk wawancara)

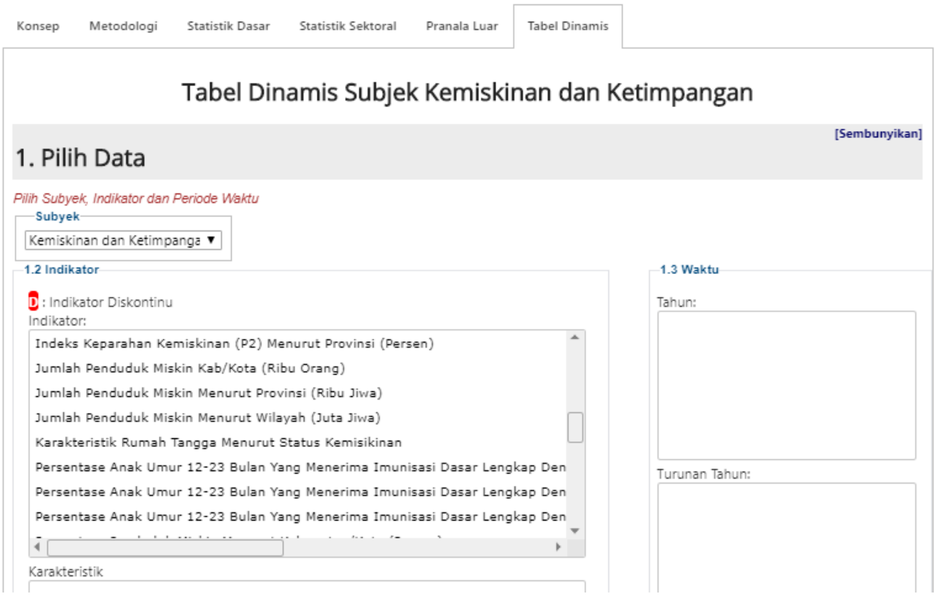
Figure 2. Home page of BPS Web Portal

The data presented are grouped into three main subjects, namely social and population, economic and trade, agriculture and mining. Data is conferred in the form of static tables and dynamic tables that allow users to define their own dimensions to be tabulated. Both can be downloaded in various formats, such as XLS files, CSV files, XML files, and PDF files. Figures 3 and 4 show how data is presented.



Kelompok Pengeluaran	Proporsi Populasi Penduduk Yang Memiliki Akses Terhadap Layanan Sanitasi Layak Dan Berkelanjutan Menurut Kelompok Pendapatan (Persen)				
	2015	2016	2017	2018	2019
Kuintil 1	41.76	46.51	48.61	50.72	62.77
Kuintil 2	51.30	57.95	58.12	59.92	70.08
Kuintil 3	58.69	64.89	65.30	68.14	76.65
Kuintil 4	68.47	74.35	73.32	74.28	82.45
Kuintil 5	81.14	85.57	85.09	86.20	89.22

Figure 3. Static Table in BPS Web Portal



Tabel Dinamis Subjek Kemiskinan dan Ketimpangan

1. Pilih Data

Pilih Subyek, Indikator dan Periode Waktu

Subyek

Kemiskinan dan Ketimpanga ▼

1.2 Indikator

D : Indikator Diskontinu

Indikator:

Indeks Keparahan Kemiskinan (P2) Menurut Provinsi (Persen)
Jumlah Penduduk Miskin Kab/Kota (Ribu Orang)
Jumlah Penduduk Miskin Menurut Provinsi (Ribu Jiwa)
Jumlah Penduduk Miskin Menurut Wilayah (Juta Jiwa)
Karakteristik Rumah Tangga Menurut Status Kemisikinan
Persentase Anak Umur 12-23 Bulan Yang Menerima Imunisasi Dasar Lengkap Den
Persentase Anak Umur 12-23 Bulan Yang Menerima Imunisasi Dasar Lengkap Den
Persentase Anak Umur 12-23 Bulan Yang Menerima Imunisasi Dasar Lengkap Den

Karakteristik

1.3 Waktu

Tahun:

Turunan Tahun:

Figure 4. Dynamic Table in BPS Web Portal

Based on the open data level suggested by Tim Berners-Lee, BPS web portal is at level 3, which allows users to download data in CSV format which is a non-proprietary open format. To go to the next level, there are steps that must be taken by BPS, which will be explained in the next section.

2.2. Linked Open Data Implementation Guide

Guidance on how to make LOD has been initiated by various parties. The Japanese Statistics Bureau published the LOD website in 2016. In his writings, (Asano. Y., et al, 2016) explains the steps of Japanese Statistics Bureau in implementing LOD. The steps are as follows:

1. Data selection: Determine the target statistical data to be converted into RDF form. In this paper, we take one example of data "number of poor people by province in 2007-2019". This data is one of the data that can be used as support for the achievement of SDGs indicators. The data used can be accessed at: https://www.bps.go.id/dynamictable/2016/01/18/1119/jumlah-penduduk-miskin-menurut-provinsi-2007-2019.html.
2. Ontology preparation: Lists items needed to declare target data as LOD. When the standard vocabulary is available, the vocabulary can be used. If it is not available, it is necessary to define vocabulary as an ontology. Some of the vocabularies that have been available are purl.org, w3.org.
3. Make a definition of dataset and data structure definition (DSD).
4. Conversion of observations into RDF. In this case, we convert a table from BPS web portal into RDF form.

3. Result:

Based on the data mentioned above, the table consists of four (4) dimensions, namely province, administrative area categories (urban / rural), year, and survey period (annual / semester). While the measure is population number, in this case the number of poor people. These dimensions and measures become basis of vocabulary or ontology development. For sampled data, there is a vocabulary that has been provided by external BPS and can be used immediately. However, it still needs to define some vocabularies internally, because of the different concepts and definitions used. Table 1 shows the list of external vocabularies used, and table 2 shows the internal vocabularies that are defined.

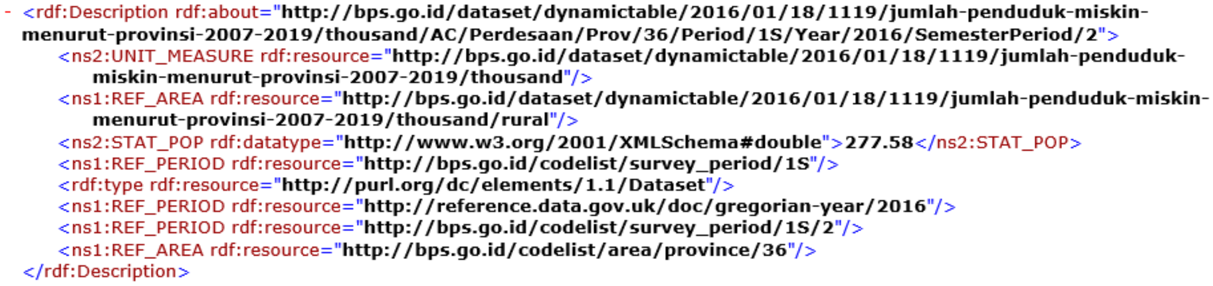
Table 1. Standard Vocabulary from External

URI
http://purl.org/linked-data/sdmx/2009/metadata#STAT_POP
http://purl.org/linked-data/sdmx/2009/metadata#OBS_VALUE
http://purl.org/linked-data/sdmx/2009/dimension#REF_AREA
http://purl.org/linked-data/sdmx/2009/metadata#UNIT_MEASURE
http://purl.org/linked-data/sdmx/2009/dimension#REF_PERIOD
http://reference.data.gov.uk/doc/gregorian-year/2007 - 2019

Table 2. Internal Defined Vocabulary

URI
http://bps.go.id/codelist/area/province/
http://bps.go.id/codelist/survey_period/
http://bps.go.id/dataset/dynamictable/2016/01/18/1119/jumlah-penduduk-miskin-menurut-provinsi-2007-2019/thousand

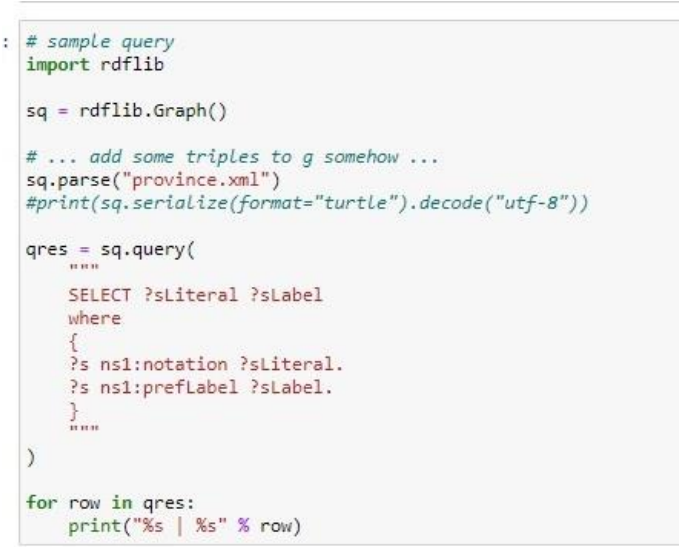
Figure 5 is a piece of RDF results from the sampled table. It shows the value of one cell in the table of data.



```
- <rdf:Description rdf:about="http://bps.go.id/dataset/dynamictable/2016/01/18/1119/jumlah-penduduk-miskin-
  menurut-provinsi-2007-2019/thousand/AC/Perdesaan/Prov/36/Period/1S/Year/2016/SemesterPeriod/2">
     <ns2:UNIT_MEASURE rdf:resource="http://bps.go.id/dataset/dynamictable/2016/01/18/1119/jumlah-penduduk-
        miskin-menurut-provinsi-2007-2019/thousand"/>
     <ns1:REF_AREA rdf:resource="http://bps.go.id/dataset/dynamictable/2016/01/18/1119/jumlah-penduduk-miskin-
        menurut-provinsi-2007-2019/thousand/rural"/>
     <ns2:STAT_POP rdf:datatype="http://www.w3.org/2001/XMLSchema#double">277.58</ns2:STAT_POP>
     <ns1:REF_PERIOD rdf:resource="http://bps.go.id/codelist/survey_period/1S"/>
     <rdf:type rdf:resource="http://purl.org/dc/elements/1.1/Dataset"/>
     <ns1:REF_PERIOD rdf:resource="http://reference.data.gov.uk/doc/gregorian-year/2016"/>
     <ns1:REF_PERIOD rdf:resource="http://bps.go.id/codelist/survey_period/1S/2"/>
     <ns1:REF_AREA rdf:resource="http://bps.go.id/codelist/area/province/36"/>
  </rdf:Description>
```

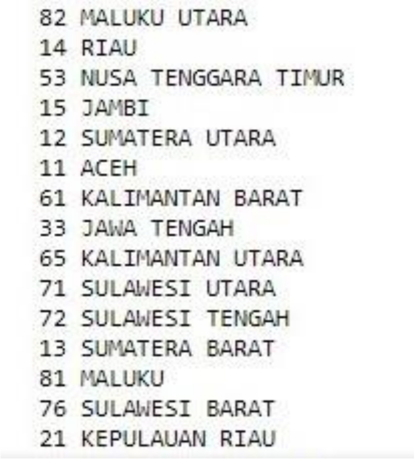
Figure 5. RDF Result Example

To retrieve information stored in RDF files, we use SPARQL as a query language. Figure 6 shows an example of SPARQL used to get a list of provinces code and name in the data. The query result is shown in figure 7.



```
# sample query
import rdflib

sq = rdflib.Graph()

# ... add some triples to g somehow ...
sq.parse("province.xml")
#print(sq.serialize(format="turtle").decode("utf-8"))

qres = sq.query(
    """
    SELECT ?sLiteral ?sLabel
    where
    {
    ?s ns1:notation ?sLiteral.
    ?s ns1:prefLabel ?sLabel.
    }
    """
)

for row in qres:
    print("%s | %s" % row)
```

Figure 6. SPARQL Example



```
82 MALUKU UTARA
14 RIAU
53 NUSA TENGGARA TIMUR
15 JAMBI
12 SUMATERA UTARA
11 ACEH
61 KALIMANTAN BARAT
33 JAWA TENGAH
65 KALIMANTAN UTARA
71 SULAWESI UTARA
72 SULAWESI TENGAH
13 SUMATERA BARAT
81 MALUKU
76 SULAWESI BARAT
21 KEPULAUAN RIAU
```

Figure 7. SPARQL Result

4. Discussion, Conclusion and Recommendations:

The application of LOD to BPS data is basically possible. It can be seen from our sampled data that has a high complexity where the data has four dimensions and can be converted into a standard structure (RDF). In the first step, BPS can determine some data that are prioritized for the initial LOD application, for example SDGs indicator data. So that, data comparability and evaluation between NSOs can be done more easily. However, one of the complexities of process in implementation LOD is defining vocabulary or ontology development. BPS with other NSOs can have different concepts or code lists, so each NSO needs to build its own vocabulary. Code lists standardization is very important to facilitate the integration of statistical data from different sources. The construction of ontology is also strongly influenced by the dimensions and measures available in the cube data. Multiple measures will make the ontology development become more complex. Another challenge in LOD implementation is building a user-friendly interface for users, because not all users are familiar with the use of queries.

References:

Asano, Y., et al (2016). Publication of Statistical Linked Open Data in Japan. 4th International Workshop on Semantic Statistics. http://ceur-ws.org/Vol-1654/article-01.pdf

Auer, S. R., et al. (2007). DBpedia: A Nucleus for a Web of Open Data. The Semantic Web. Lecture Notes in Computer Science. 4825. p. 722. doi:10.1007/978-3-540-76298-0_52. ISBN 978-3-540-76297-3.

Bauer, F., & Kaltenböck, M. (2012). Linked Open Data: The Essentials. Edition mono/monochrom, Vienna, Austria.

Blaney, J. (2017). Introduction to the Principles of Linked Open Data. https://programminghistorian.org/en/lessons/intro-to-linked-data.

Tim Berners-Lee. 5 Star Open Data. https://5stardata.info/en/.