Indonesian Journal of Electrical Engineering and Computer Science
Vol. 16, No. 3, December 2019, pp. 1521~1529
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v16.i3.pp1521-1529

Generating RDF resources from web open data portals

Khalid S. Aloufi
Computer Engineering Department, College of Computer Science and Engineering, Taibah University, Saudi Arabia

Article Info

Article history:

Received Apr 1, 2019
Revised Jun 28, 2019
Accepted Jul 22, 2019

Keywords:

Data sharing
Linked data
Metadata
Open data
Semantic web

ABSTRACT

Open data are available from various private and public institutions in different resource formats. There are already great number of open data that are published using open data portals, where datasets and resources are mainly presented in tabular or sheet formats. However, such formats have some barriers with application developments and web standards. One of the web recommenced standards for semantic web application is RDF. There are various research efforts have been focused on presenting open data in RDF formats. However, no framework has transformed tabular open data into RDFs considering the HTML tags and properties of the resources and datasets. Therefore, a methodology is required to generate RDF resources from this type of open data resources. This methodology applies data transformations of open data from a tabular format to RDF files for the Saudi Open Data Portal. The methodology successfully transforms open data resources in sheet format into RDF resources. Recommendations and future work are given to enhance the development of building open data.

Corresponding Author:

Khalid S. Aloufi,
Computer Engineering Department,
College of Computer Science and Engineering, Taibah University, Saudi Arabia.
Email: koufi@taibahu.edu.sa

1. INTRODUCTION

With the global initiatives on publishing data on the Internet, data are being provided by different governments and organizations around the world. The data generated by the Internet of Things (IoT) are transmitted over the Internet and shared between applications [1]. Governments, such as the UK, Taiwan and Saudi Arabia, now publish and regularly update their open data [2-4]. Taiwan is ranked number eleven in 2014; however, it was ranked number one in 2015 and 2019, with a score of 78% according to the statistics published by Global Open Data [2, 5]. In addition, the UK was ranked first in 2014 and second in 2015, with a score of 76% [3]. The list presents 122 countries, where the 122th country scored 3% [5].

Saudi open data were ranked 103rd in 2015, down from 74th in 2014, with a score of 15% [5, 6]. Australia are moving from the 9th in 2013, 5th in 2015 to the in 2nd position in 2019 [5]. For example, some open data portals [ODP] are not listed, which affects the score of the country being analysed. Data Portals list ODP worldwide [7].

Currently, Open Data on different government ODP are presented in tabular or other formats. In general, statistics and URLs are available in the ODP. However, there are neither measure the quality of the available open data nor a method to use the data in applications. Different studies provide applications of using data [8]. However, the data are not presented in structured open format and easily presented as open data [9]. Using semantic web technologies, there is an API to publish the data using the ODP, however, there is not an API on how to use the datasets or resources published by ODP. One of the goals in making data openly available is to make the data ready for application and linking with other data,linked data. ODP present Data in arbitrary formats are not ready for web applications. Therefore, data should be published in the semantic web standard: The Resource Description Framework [RDF] format. This research concerns

transforming already published tabular data into the RDF format. For example, Table 1 shows an information that could be represented in an RDF, shown in Figure 1.

Table 1. Maximum Temperatures in 2012 for cities in Saudi Arabia

City	Month	
	December	June
Turaif	19.0	42.0
Arar	23.8	46.0
Damam	32.5	47.6
Riyadh	30.0	46.7

The Table 1 consists of the header on the top and right and the data in the body. To represent the data in RDF triples, the data should be transformed such that the subject represents the row header, the predicate represents the column header, and the object represents the cell value. However, this task is not trivial because of the different table designs found in ODP, as will be shown later.

```
    xmlns:sa="http://data.gov.sa/dataset#" ...
  <rdf:Description rdf:about="http://data. gov.sa/dataset#Dammam">
    <sa:December
    rdf:datatype="http://www.w3.org/2001/XMLSchema#double">32.5</sa:December>
    <sa:June rdf:datatype="http://www.w3.org/2001/XMLSchema#double">47.6</sa:June>
    </rdf:Description>
...
```

Figure 1. RDF/XML of part of the data shown in Table 1

ODP present *** datasets without considering the five stars open data rankings, as recommended by Tim Berners-Lee [7]. The data stars are

- * - when the data are online,
- ** - when the data are readable by certain software and are structured,
- *** - when the data are in a non-proprietary open format,
- **** - when the data in RDF and include URIs,
- and ***** - when the data are linked to other data.

Data are *** * data when they are available online and structured, in an open format and include URIs. Five-star open data are four-star open data linked to other sources, called Linked Open Data [LOD] [10]. During the development of RDF data, ranking and guidelines should be considered. The four guidelines for linked data are as follows [11, 12]:

- URI is used to name objects or any thing.
- HTTP URI is used to locate resources.
- RDF and SPARQL are used to add and process information about the URIs.
- URIs are added to other resources to create links between resources.

This research presents a framework for transforming tabular data into RDF. The framework is developed and tested on the Saudi ODP [4]. During the development of the framework, different table designs were considered. This methodology transforms two-star open data, in xls format, into four-star open data, which can be transformed to five star open data, and that is left for future work, however, making four star data into five star is well presented in the research and commercial fields. This research applied data transformations from tabular formats presented in xls sheet files to RDF.

In addition, this research mainly concerns the presentation and arrangement of open data resources that are presented in tabular or sheet formats such as xls files. Open data has taken a great attention and still one of the hot research topics and commercial products. Different frameworks have been developed to generate high-quality open data in the RDF resource format [13-18].

After this introduction, a detail about ODP is presented in section 2 with a use case selected as the Saudi ODP. Then, section 3 presents the proposed framework. Then, section 4 present the generated RDF resources. Finally, the paper close with the conclusion section 5 with some emphasis n future work as well.

2. METHODOLOGY

In this section, detailed explanation about open data portal is presented. The study is applied in the Saudi open data portal. The framework, presented in section 3, is applied to the Saudi ODP [19]. The generated RDF resources have been verified successfully by the RDF validation [20]. Details of the Saudi open data are available at [20]. In summary, there are 319 datasets and 6,744 resources in Saudi ODP, where the datasets are displayed according to the following different filters which are by group as shown in Table 2, by publisher, by resource format as shown in Table 3 and by license type. When selecting one group, a dataset will show the available resources. Each web page in the portal contains information or metadata about its contents. In ODP, metadata are published to describe the datasets under a specified standard format such as DCAT [21]. The portal uses the recommended DKAN framework to build the ODP [22]. The main page of the data portal does not contain any semantic vocabularies.

Table 2. Filtered by Group

Group	datasets	Resources	Triples
Arab Gulf Cooperation Council (GCC)	12	241	261212
Industry	12	126	85087
Labor Market	14	248	1179896
Transport and Communications	29	568	1898427
Energy and Water (GCC)	13	92	260908
Health	50	68	902514
Accounts Financial Monetary Affairs and Industry	28	318	1122189
Population and Housing	20	291	615187
Agriculture and Fishing	27	783	1345461
Social Services	50	816	1799234
Education and Training	16	358	1724813
Trade (internal and external)	14	249	829987
Social Insurance	5	239	674968
Prices and Indices	8	84	31477
Weather Conditions	12	201	710824
Totals	310	4682	13442184

Table 3. Filtered by resource format

Type	Number
xls	(287)
xlsx	(226)
jpg	(35)
xml	(5)
xlb	(2)
jpeg	(1)

2.1. Open Data Portal

When applying the algorithm of the framework to the resources in xls format, some difficulties were presented. These difficulties results in some recommendations for designing open data as will be shown later. The following difficulties were found in processing the open data resources:

- Hidden URLs of the resource files, which make programming direct online access not immediately possible.
- Some data have certain symbols or naming abbreviations that are not clarified for later data linking.
- Open data include repeated data such as when having data about water consumption from 2003 to 2010. Data published later includes data from 2003 to 2013; two years of data are added, and the remaining data are duplicates.
- Old data are not yet published such as data prior to 2003.
- Some numbers are written as a string, for example, "1" is written "'1" when most of the data are integers.
- There are no standards for publishing the data.
- Using different sheets for one xls resource with internal references between cells in different sheets.
- Different table structures of the xls resources.
- There are different resource formats; however, this research only considers xls.
- The publishers did not have a standard for the data formats, structure or resource names.
- Duplicate links to the datasets and resources.
- Different dates for the resources compared to their title.
- Empty spaces in some tables.

- Hidden rows and columns.
- Some resources are empty.
- Some resources are partially filled.
- Data are not connected to other sources when building the data.
- Certain characters should not be used, such as ILLEGAL_CHARACTER in FRAGMENT: the character violates the grammar rules for URIsIRIs.
- No API to simplify access to the data in different formats.
- There were different ways of presenting the resources.
- Guidelines are needed to show how to use and build the open data.
- Provide a SPARQL end point to simplify access
- Open data should be in one of the formats defined in one of the W3C standards for data.
- After providing the data from the source, the data must be tagged with data to define them using a defined vocabulary such as The Data Catalog Vocabulary [DCAT] [21].
- Different properties of the data, such as the title of the dataset and the issue date, should be added to the resource table or RDF.

Algorithm 1 Creating RDF resources from the open portal

Input: XLS resources from ODP
Output: RDF resources

- Open the group URL
- Open the dataset page
- List all the properties of the dataset
- Open the resource page
- List all the properties of each resource • If the resource is in XLS format, then
  - Define the first upper-right cell of the data – Define the upper header
  - consider any header cell as a sub-property of its upper header cell.
  - Define the right header
  - Define the left header
  - link the right header with the left header with the property owl:sameAs.
  - Create the RDF triples for each cell of the data such that
    - Subject represents the row header
    - Predicate represents the column header
    - Object represents the cell value.
- Create the RDF resource from the properties collected from the dataset as well as the resource and the data.
- Publish the data resources in the RDF format.

2.2. The proposed framework

The framework includes determined steps to gather the data from the portal. The proposed framework is shown in Figure 2. First, the main web page of the open data is accessed. The main web page of the portal classifies open data by group, publisher, resource format and license. The Group URL is selected to access the data. Then, the framework starts with traversing through the ODP web pages, dataset web pages and resource web pages.

The system navigates through the groups one by one and creates the RDF information for each group. Then, it navigates through the datasets in each group one by one and creates the RDF information for each dataset. After that, The properties of the datasets and resources are collected from the web pages and presented in RDF format,then navigates through the resources in each dataset one by one. The prefixes, listed in Table 4, are presented as meta data on the HTML web pages of the resources and the datasets. The prefixes are used as the main top part of the generated RDF resources.

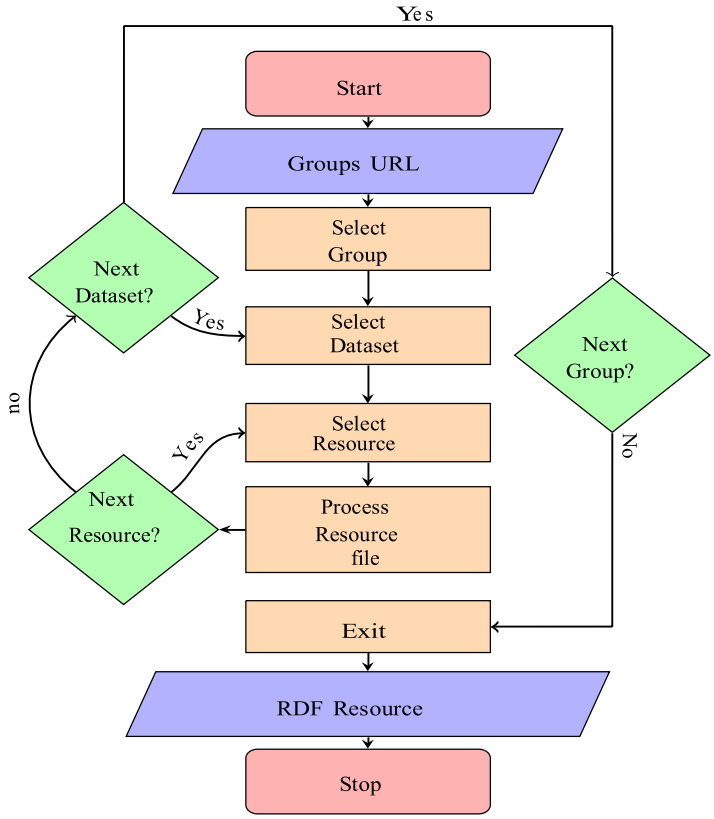

Figure 2. Flowchart of the system

Table 4. Vocabularies

xmlns:content="http://purl.org/rss/1.0/modules/content/"
xmlns:dc="http://purl.org/dc/terms/" [23]
xmlns:foaf="http://xmlns.com/foaf/0.1/"
xmlns:og="http://ogp.me/ns#"
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
xmlns:sioc="http://rdfs.org/sioc/ns#"
xmlns:sioct="http://rdfs.org/sioc/types#"
xmlns:skos="http://www.w3.org/2004/02/skos/core#"
xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
xmlns:owl="http://www.w3.org/2002/07/owl#"
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:rss="http://purl.org/rss/1.0/"
xmlns:site=http://data.gov.sa/ns#
xmlns:dcat="http://www.w3.org/ns/dcat#" [21]
dcterms, not included but used in the html properties [23].

Then, the tabular data of each resource presented in xls are transformed into RDF. For an xls resource, the framework defines the first data cell, from which it defines the table structure. This is very important steps and it is used because of the different structures of the several of resources available. The framework assumes that the tabular data cells are integers, whereas other cells, such as header cells, are characters.

In the resource, for each data cell, the subject represents the row header, the predicate represents the column header, and the object represents the cell value. The final RDF resource for each resource is created with the RDF information of the corresponding resource, dataset and group. Finally, the RDF of each resource is created in RDF format and includes the prefixes, the dataset properties, the resource properties and the data.

The RDF resource can then be published on the web page of the resources and queried in SPARQL. The framework stops when completing the cycles for all datasets and resources in the portal. The framework is composed of the steps shown in Algorithm 1; however, the algorithm shows some more details as follows.

All the steps of the algorithm are clearly shown and follows the framework steps. When the table has top, left and right headers, the top header is the predicate of the data cells, and the right header is linked with the left header with the property owl: sameAs. Each step is performed for each cell in the table. Each cell of the top header is the sub-property of its upper cell. In addition, the data cell of the body is an object. The top header cell is the subject, and the right or left cell is the predicate. In this research, the right cell is selected as the predicate because most of the resources are using Arabic as the main language of the data.

The algorithm generates the RDF from the data after defining the RDF graph of any row for each resource, which yields the final graph as will be shown later in section 4. Each xls resource has a specific RDF graph, as each row is identical in the structure. The RDF graph of each row is used by the algorithm to transform the tabular data into the RDF format. In this research, the graphs of all rows are identical. As mentioned by [24], each row G_i generates a map G_r, and the whole table yields the final map:

$$G = \bigcup_{i=1}^{row_counts} G_i$$

$$\boldsymbol{G_i = map(rowCells[], i)}.$$

3. RESULTS AND DISCUSSION: THE RDF RESOURCE

After converting the xls resource into the RDF format according to the algorithm mentioned earlier in Section 3, the RDF file is then created from the dataset, the resource web pages and the resource sheet file. The experiment included processing of 4682 files, with a total size of 706.4 MB. The average number of triples in one resource is 3482 triples. The average size of an RDF resource file is 150 KB. One of the examples of the tables is shown in Table 5 and part of its generated RDf is shown in Figure 3 and the RDF graph is shown in Figure 4 and 5. The properties of the datasets are summarized in Table 6. Each dataset web page lists the available distribution and the resource formats. Usually, these files are generated in different formats. Table 7 shows the properties of a resource found in the resource web page such as the format and the update date. After creating the RDF resource, the data could be linked to other linked data. All properties of the dataset and the resource form the RDF Graph. Because the tables in the open data have multi-level headers, the lower header is considered as a sub-property of the upper header, as shown in Figure 7.

Figure 6 show examples of a SPARQL queries over the generated RDF resources and Table 8 show the query results. The data can be used to generate representations of the data on any web page or document from the data.

Table 5. Maximum Temperature of PME MET Stations: 2012 [19]

Weather Conditions الاحوال الطبيعية

بيانات الحرارة العظمى لمحطات الرصد الجوى التابعة للرئاسة العامة للأرصاد و حماية البيئة لعام 2012

MAXIMUM TEMPERATURE OF PME MET STATIONS:2012

(DEGREE CENTIGRADE) وحدة القياس الدرجة المئويه

Table1-2 جدول1-2

Month / station	ديسمبر December	نوفمبر November	أكتوبر October	سبتمبر September	أغسطس August	يوليو July	يونيو June	مايو May	أبريل April	مارس March	فبراير February	يناير January	الشهر / المحطة
Turaif	19	30	35	39	41	43	42	36	33	25	22	17	طريف
Arar	24	34	39	42	46	48	46	40	37	30	27	23	عرعر
Guriat	23	33	37	40	41	46	45	40	35	28	26	18	القريات

```
<rdf:RDF ...
<rdf:Description rdf:about="http://data.gov.sa/dataset/temperature/resource/2fbaac10-b87a-4a8-b792-
      8aaaa1afb2ef# March">
<rdf:type rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Property"/> ...
<saResource:January rdf:datatype="http:// www.w3.org/2001/XMLSchema#double">29.4</saResource:January> ...
      <saResource:December
      rdf:datatype="http://www.w3.org/2001/XMLSchema#double">32.4</saResource:December>
</rdf:Description>
...
</rdf:RDF>
```

Figure 3. RDF/XML of the city Yenbo data shown in Table 5

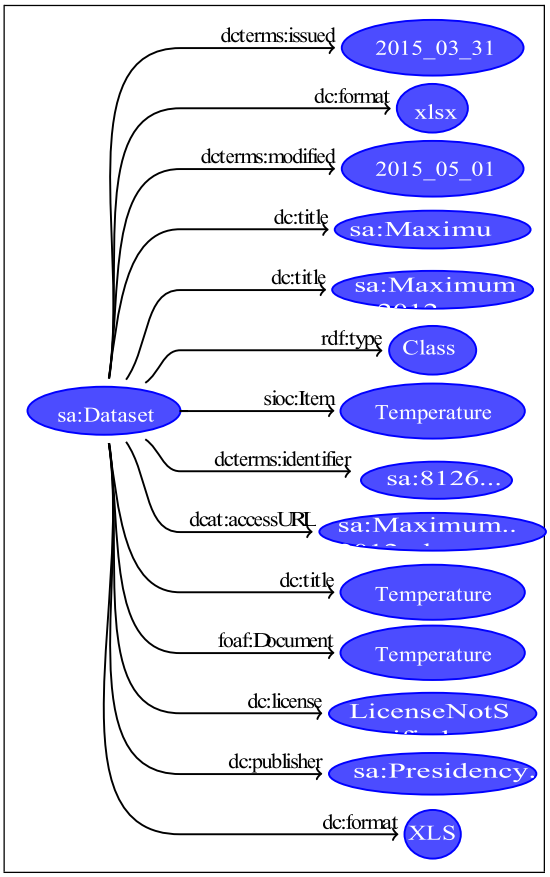

Figure 4. Dataset RDF Graph [19]

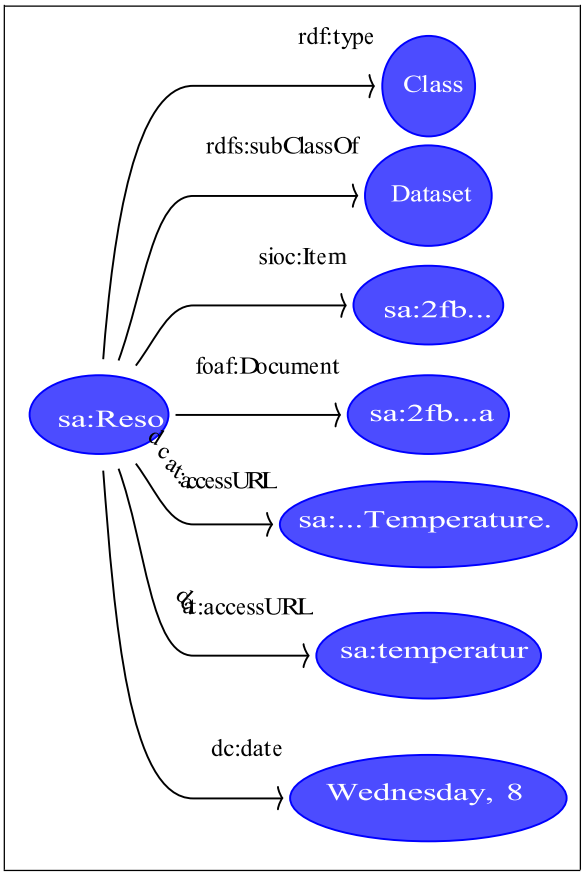

Figure 5. Resource RDF graph of [19]

Table 7. Properties of the resource

sioc:Item	the dataset URI
foaf:Document	the dataset URI
dc:title	the title of the dataset
dc:group	the group of the dataset.
dc:publisher	the publisher of the dataset
dcterms:modified	the last modified date
dcterms:issued	the Release Date
dcterms:identifier	serial number
dc:license	License
dcat:distribution	available resources format
dcat:accessURL	access URL
dc:format	resource or distribution format

Table 8. Yanbu Average Temperature

Month	Average Temperature
saT:October	"40.6"8sd:double
saT:May	"44.5"8sd:double
saT:April	"36.8"8sd:double
saT:November	"37.0"8sd:double
saT:June	"47.2"8sd:double
saT:December	"32.4"8sd:double
saT:January	"29.4"8sd:double
saT:July	"46.0"8sd:double
saT:September	"43.0"8sd:double
saT:August	"48.4"8sd:double
saT:March	"33.6"8sd:double
saT:February	"32.9"8sd:double

```
PREFIX sa:<http://data.gov.sa/dataset#>
PREFIX xsd:<http://www.w3.org/2001/XMLSchema#>
PREFIX rdf:<http://www.w3.org/1999/02/22rdf-syntax-ns#>
PREFIX rdfs:<http://www.w3.org/2000/01/rdfschema#>
PREFIX saT:<http://data.gov.sa/dataset/ temperature/resource/2fbaac10-b87a-4aa8-b792-8aaaa1afb2ef#>
PREFIX owl:http://www.w3.org/2002/07/owl#

SELECT ?p ?o
WHERE {saT:Yanbu ?p ?o.}
```

Figure 6. SPARQL QUERY for a specific column

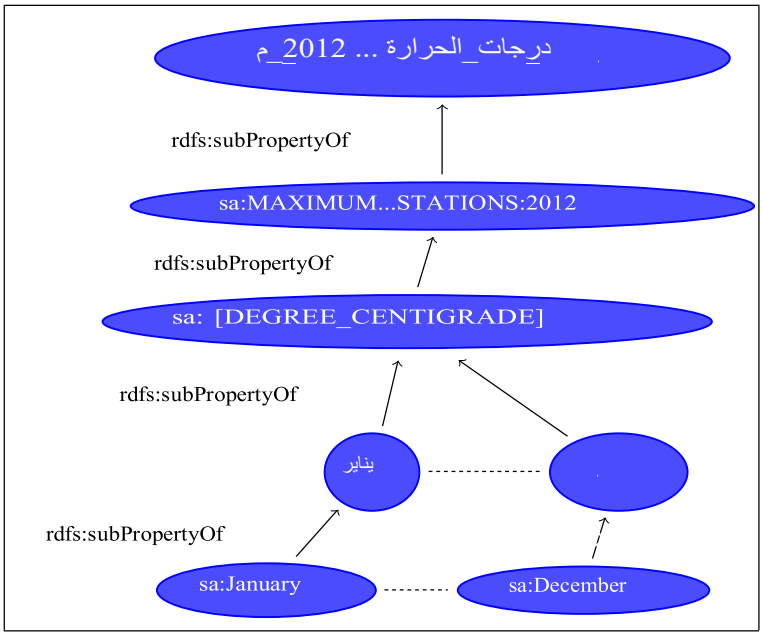

Figure 7. Property and sub-property RDF graphs of [23]

4. CONCLUSION

This research presents a framework to generate RDF resources from xls resources. The framework is applied to the Saudi ODP. Recommendations are presented on developing open data. The application of the framework is shown for one xls resource. A detailed explanation of the output RDF is presented. Open data will be important aspect of future web applications when it gains the confidence of the developer community. The sources of the open data are responsible for access and the accuracy of the data. Usually, the community continues to use open data when the data are found to be useful and accurate.

Conversely, the community will stop using the open data when the data are found to be difficult to access or inaccurate. Future work includes the development of a framework to generate the data in 5 star linked data in RDF format, publish and provide access to the data. In addition, future work will investigate the usability of these data and present methods for applying machine learning for decision making enhancing applications, such as for streaming open data. Finally, the framework could be extended to generate data from various available resources such a tables, images, and PDFs.

REFERENCES

[1] F. S. Adnan, et al., "Testbed versus simulation approach on RF communication with AAβ asymmetric encryption scheme on internet of things devices," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 14, pp. 353-359, 2019.
[2] T. O. D. Index, "Tracking the state of government open data," 2016. Available: http://index.okfn.org/place/taiwan/.
[3] U. O. Data, "UK open data," 2016. Available: http://data.gov.uk/.
[4] S. N. e Government Portal, "Saudi government open data portal," 2015. Available: http://data.gov.sa/.
[5] O. D. Index, "Tracking the state of government open data," 2016. Available: http://index.okfn.org/place/.
[6] S. A. O. Data, "Saudi arabia open data," 2016. Available: http://index.okfn.org/place/saudi-arabia/.
[7] D. Portals, "Data portals a comprehensive list of open data portals from around the world," 2016. Available: http://dataportals.org.
[8] H. Ahuja and Sivakumar R., "Implementation of FOAF, AIISO and DOAP ontologies for creating an academic community network using semantic frameworks," *International Journal of Electrical and Computer Engineering*, vol. 9, pp. 4302-4310, 2019.
[9] T. Adiono, et al., "Curtain Control Systems Development on Mesh Wireless Network of the Smart Home," *Bulletin of Electrical Engineering and Informatics (BEEI)*, vol. 7, 2018.
[10] T. Berners-Lee, "5 star open data," 2016. Available: http://5stardata.info/en.
[11] T. Berners-Lee, "Linked data," 2015. Available: http://www.w3.org/DesignIssues/LinkedData.html.
[12] L. Han, et al., "RDF123: from spreadsheets to RDF," *The Semantic Web - ISWC 2008, 7th International Semantic Web Conference, ISWC 2008, Karlsruhe, Germany, Proceedings, Lecture Notes in Computer Science, Springer*, vol. 5318, pp. 451-466, 2008.
[13] W. Beek, et al., "LOD laundromat: A uniform way of publishing other people's dirty data," *The Semantic Web - ISWC 2014 - 13th International Semantic Web Conference, Riva del Garda, Italy, Proceedings, Lecture Notes in Computer Science, Springer*, vol. 8796, pp. 213-228, 2014.
[14] I. Ermilov, et al., "User-driven semantic mapping of tabular data," *ISEMANTICS 2013 - 9th International Conference on Semantic Systems, ISEM '13, Graz, Austria, ACM*, pp. 105-112, 2013.
[15] I. Ermilov, et al., "User-driven semantic mapping of tabular data," *Proceedings of the 9th International Conference on Semantic Systems, I-SEMANTICS '13, ACM, New York, NY, USA*, pp. 105-112, 2013.
[16] I. Ermilov, et al., "Csv2rdf: User-driven csv to rdf mass conversion framework," *Proceedings of the ISEM '13, Graz, Austria*, 2013. Available: http://svn.aksw.org/papers/2013/ISemantics_CSV2RDF/public.pdf.
[17] T. Knap, "Increasing quality of austrian open data by linking them to linked data sources: Lessons learned," *Joint Proceedings of the 2nd Workshop on Managing the Evolution and Preservation of the DataWeb [MEPDaW 2016] and the 3rd Workshop on Linked Data Quality [LDQ 2016] co-located with 13th European Semantic Web Conference [ESWC 2016], Heraklion, Crete, Greece, CEUR Workshop Proceedings, CEUR-WS.org*, vol. 1585, pp. 52-61, 2016.
[18] D. U. Board, "Dcmi metadata terms [dcterms]," 2015. Available: http://dublincore.org/documents/dcmi-terms/.
[19] W. R. Validation, "World wide web consortium [w3c]," 2016.
[20] K. Aloufi, "Rdf resources of saudi open data," *Proceedings of the 2015 International Conference on Recent Advances in Computer Systems, racs '15, Atlantis Press*, 2015. Available: http://uohapp.uoh.edu.sa/racs/.
[21] S. N. e Government Portal, "National portal," 2015. Available: http://www.saudi.gov.sa/wps/portal/.
[22] R. Kamala and R. J. Thangaiah, "An Improved Hybrid Feature Selection Method for Huge Dimensional Datasets," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 8, pp. 77, 2019.
[23] F. Maali, et al., "Data catalog vocabulary [dcat]," 2015. Available: http://www.w3.org/TR/vocab-dcat.
[24] C. Bizer, et al., "Linked data - the story so far," *Int. J. SemanticWeb Inf. Syst.*, vol. 5, pp. 1-22, 2009.

BIOGRAPHIES OF AUTHORS

Khalid Aloufi is an associate professor in the Department of Computer Engineering, Taibah University, Madinah, Saudi Arabia. He received his Ph.D. and M.Sc. degrees in Computing from Bradford University, UK, in 2002 and in 2006 respectively. His B.Sc. degree in computer engineering was received in 1999 from King Fahd University of Petroleum and Minerals [KFUPM], Saudi Arabia. From 2002 to 2006, he was part of the networks and performance engineering research group at Bradford University. Since 2013, Dr. Aloufi has been the dean of the College of Computer Science and Engineering at Taibah University, Saudi Arabia.