Linking Open Drug Data: The Arabic Dataset

Guma Lakshen*, Valentina Janev**, Sanja Vraneš***
*School of Electrical Engineering, University of Belgrade, Serbia.
**, ***Mihajlo Pupin Institute, University of Belgrade, Serbia.
*E-mail: Jlackshen65@yahoo.com
**E-mail: valentina.janev@institutepupin.com
***E-mail: sanja.vranes@institutepupin.com

Abstract— **Linked Open Data illustrates the concept that provides an optimum solution for information and dissemination of data, through the representation of the data in an open machine readable format and to interlink it from diverse repositories to enable diverse usage scenarios for both humans and machines.**

The pharmaceutical/drug industry was among the first that validated the applicability of the approach for interlinking and publishing open linked data. Yet, open issues arose clearly when trying to apply the approach to datasets coded in languages other than English, for instance, in Arabic languages. Author's objective is to examine in detail the requirements specification process for building Linked Data application taking into consideration the possibility of reusing recently published datasets and tools. Main conclusions derived from this study are that making drug datasets accessible and publish it in an open manner in linkable format adds great value by integration to other notable datasets. Author's main contribution is the enhancement of Arabic knowledge graph based on drug data from selected Arabic countries and the novel methodology for building Linked Data applications.

Keywords: Linked data, drugs application, datasets, and methodology.

I. INTRODUCTION

Data volume is growing at a tremendous rate. International Data Corporation[1] estimates that by the end of this year as much as 33 zettabytes of useful data will exist in the Web, while it is also expected to reach 175 zettabytes by 2025. Innovative technology offers new methodologies of retrieving value from the huge big data available all around the world. Effective information management and knowledge extraction are now considered as a key competitive advantage. The adoption of big data technologies started to be realized by organizations as a must and an imperative necessity to survive and gain competitive advantage.

Linked Data principles [1] adopted data publishers from researchers and domains alike led to the creation of the Linked Open Data (LOD) Cloud, where a vast collection of interlinked data published on and accessible via the existing Web infrastructure, which in turn open possibilities for creating new methodologies for transforming, linking and publishing Linked Data [2].

Most organization and domains such as health, medical/pharmaceutical manufacturing, education, energy, and industry, etc.; are building their core business and services on their ability to collect and analyze information in order to extract business knowledge and insight. The benefits of sharing and Linked Open Data across domains and industry are becoming obvious in all domains.

The LOD cloud datasets http://lod-cloud.net/ have increased from 12 datasets in 2007 to 1,239 with 16,147 links (as of March 2019)[2].

The pharmaceutical/drug industry was leading other domains in expressing interest in validating the approach for publishing and integrating open data. Linked Open Drug Data LODD endpoint was created in 2011, https://www.w3.org/wiki/HCLSIG/LODD, which is a set of linked datasets related to Drug Discovery [3]. It includes data from several datasets including DrugBank, DailyMed, LinkedCT, SIDER, ClinicalTrials.gov, RxNorm, and NCBI Entrez Gene, a detailed comparison of the LODD datasets can be accessed at https://www.w3.org/wiki/HCLSIG/LODD/Data, notably, this page was last updated on 28th December 2012. Later, in 2014, the 3rd release of Bio2RDF(http://bio2rdf.org/ or https://github.com/bio2rdf/) was published as the largest network of Linked Data for the Life Sciences (35 datasets).

The aim of this paper is to explore a methodology development of the Linked Drug Data Application in the Arabic region, and emphasizing on the advantages and issues of using the Linked Data approach in the pharmaceutical/drug industry. The datasets can be accessed through SPARQL endpoint http://aldda.b1.finki.ukim.mk/sparql, it exposes in detail a sequential steps the requirements specification process for building Linked Data application taking into consideration the possibility of reusing published datasets including the DrugBank[3] and DBpedia[4] datasets (see Table 1).

The authors have presented a methodology which shows the necessary steps to create a linked open data

[1] https://www.forbes.com/sites/tomcoughlin/2018/11/27/175-zettabytes-by-2025/#79bff1675459

[2] https://www.lod-cloud.net/ (accessed 28/04/2019).
[3] https://www.drugbank.ca/
[4] https://wiki.dbpedia.org/

from a second class data format where quality assessment is strongly revised at every stage of the transformation process to ensure the value of the data to the end user. Also, the authors showed the added value of data due to the linkage process as for the Arabic drug data.

TABLE 1.
DATASETS LINKED TO ARABIC DRUG DATASET

DataSet	Description	Link
DrugBank	DrugBank is a web-enabled database containing comprehensive molecular information about drugs, their mechanisms, their interactions, and their targets. First described in 2006, DrugBank has continued to evolve over the past 12 years in response to marked improvements to web standards and changing needs for drug research and development.	https://www.drugbank.ca/
DBpedia	DBpedia is an ongoing project designed to extract structured data from Wikipedia. It contains RDF data about 2.49 million things out of which is 218 million triples describing 2300 drugs. DBpedia is updated every three months.	https://www.DBpedia.org

The objective is to outline a methodology for publishing a set of discrete datasets in RDF data format and to examine in detail the requirements specification process for building Linked Data application taking into consideration the possibility of reusing recently published datasets such as the DrugBank and DBpedia.

The main contribution is related to the establishment of knowledge graph based on drug data from selected Arabic countries and establishes a novel methodology for building Linked Data applications that take into consideration quality issues with open linked data, in addition, SPARQL endpoints are presented to show the value added of utilizing Linked data technologies in enhancing the value of original datasets in the Arabic language.

The paper is organized as follows: Section II describes the main research motivation. Section III gives a review of literature of some related works, whereas Section IV presents in detail the required steps needed for the process of transforming the Arabic drug datasets as a linked data. The paper closes with Section V, which discuss the process and the proposed solutions.

II. Motivation

Due to the standardization and development of semantic web technologies, data being published on the Web as LOD added a tremendous value to institutions, research centers, and enterprises. One such example is the pharmaceutical/drug industry. The amount of open data on the web that increases rapidly in drug and medicine that opens new opportunities and horizons for enhancing and integrating drug knowledge on a global scale.

As far as medical data availability in the Arab region, there are only a handful of Arabic drug applications such as Webteb, Altibbi, 123esaaf, etc., which provide their services in Arabic and English, but unfortunately, their data are not open and mostly not free. The Arabic language content in the World Wide Web is less than 3%; the situation is even worse regarding Arabic open data, Arabic linked data, and Arabic drug open linked data. This limitation of Arabic content encourages the researcher to enrich the Arabic user to gain value by utilizing semantic web technologies to interlink their data with other languages e.g. English.

This motivates the authors to enable end-user from Arabic speakers to inquire about drug availability in the local database and/or on the Web of data. The end-user will benefit from the interlinking of private datasets with public data (Drugbank, DBpedia) and enrichment of local data with information from the Web. The authors seek to develop tools and answer end-user inquiries about drug availability in the datasets with public data (DrugBank, DBpedia) and enrichment of local data with information from the Web. Examples of key business queries are:

Query1: For a particular drug, retrieve relative information in Arabic language (if exists) from other identified datasets, such as DrugBank and DBpedia.; ***Query2:*** For a particular drug, retrieve equivalent drugs; and compare active ingredients, contradictions, and prices,;
Query3: For a particular drug, retrieve valuable information about equivalent other drugs with different brand/commercial name, manufacturer, strength, form, price, etc.;

Query4: For a particular drug, retrieve its reference information to highlight possible contradiction e.g. drug/drug, drug/allergy, drug/special cases (e.g. Pregnancy), etc.;

Query5: For an active ingredient retrieve advanced clinical information i.e. pharmacological action, pharmacokinetics, etc.;

Query6: Compare prices for a particular; drug, showing drug, cost, manufacturer, and country.

The authors are aiming to enrich the current literature by facilitating the publication process of institutional datasets as LOD using the five stages methodology, also demonstrating the methodology in the drugs domain, where the authors takes a repository of drug data as an input and publish it as LOD data according to the 5-star model as proposed by Tim Berners-Lee.

III. Related Works

In literature, not many papers dealt with Linked Data methodologies i.e. the process of generating, linking, publishing and using Linked Data, to name a few; (*W3C-Government Linked Data Working Group*, 2014), (*Hyland et al.*, 2011), (*Hausenblas et al.*, 2016), (*Villazón-Terrazas et al.*, 2011), (*Auer, et al*., 2012), and (*Jovanovic and Trajanov*; 2017), see table 2 below for a brief comparison.

One of the first Linked Data methodologies was developed in the European research project LOD2 (2010-

TABLE 2.
SURVEY ON LINKED DATA METHODOLOGY

Authors/ Organization	Title / Steps	
W3C Government Linked Data Working Group (2014) [5]	***Best Practices for Publishing Linked Data*:**	
	(1) Prepare stakeholders, (2) Select a dataset, (3) Model the data, (4) Specify an appropriate license, (5) Good URIs for linked data, (6) Use standard vocabularies,	*Initialization*
	(7) Convert data, (8) Provide machine access to data,	*Innovation*
	(9) Announce new data sets, (10) Recognize the social contract	*Validation &Maintenance*
Hyland et al. (2011) [6]	***A Cookbook for Publishing Linked Government Data on the Web*:**	
	(1) Identify, (2) Model, (3) Name, (4) Describe,	*Initialization*
	(5) Convert, (6) Publish,	*Innovation*
	(7) Maintain	*Validation &Maintenance*
Hausenblas et al. (2016) [7]	***Linked Data Life Cycles*:**	
	(1) Data awareness, (2) Modeling,	*Initialization*
	(3) Publishing, (4) Discovery, (5) Integration,	*Innovation*
	(6) Use-cases	*Validation &Maintenance*
Villazón-Terrazas et al. (2011) [3]	***Guidelines for Publishing Government Linked Data*:**	
	(1) Specify, (2) Model,	*Initialization*
	(3) Generate, (4) Publish,	*Innovation*
	(5) Exploit	*Validation &Maintenance*
Auer, et all. (2012) [1]	***Managing the Life-Cycle of Linked Data with the LOD2 Stack*:**	
	(1) Extraction,	*Initialization*
	(2) Storage, (3) Authoring, (4) Interlinking, (5) Classification,	*Innovation*
	(6) Quality, (7) Evolution/Repair, (8) Search/ Browsing/ Exploration	*Validation &Maintenance*
Jovanovik and Trajanov (2017) [4]	***Methodological guidelines for consolidating drug data*:**	
	(1) Domain and Data Knowledge, (2) Data Modeling and Alignment,	*Initialization*
	(3) Transformation into 5-star Linked Data, (4) Publishing the Linked Data Dataset on the Web,	*Innovation*
	(5) Use-cases, Applications and Services	*Validation &Maintenance*

2014) that was mainly dedicated to the publishing process, i.e. opening the data in a machine-readable format and establishing the prerequisite tools and technologies for interlinking and integration of heterogeneous data sources in general [1].

In [4], Jovanovik and Trajanov concluded that, "*the LOD2 methodology which provides software tools for the denoted steps still misses some key elements of the Linked Data lifecycle, such as the data modeling, the definition of the URI format for the entities and the ways of publishing the generated dataset"*. They also stated, "*The LOD2 tools are general, and cannot be applied in a specific domain without further work and domain knowledge*."

Therefore, they proposed a new Linked Data methodology with a focus on reuse. It provides guidelines to data publishers on defining reusable components in the form of tools and schemas/services for the given domain (i.e. drug management).

IV. LINKING ARABIC DRUG DATASETS

As a use case scenario, the authors selected four drug data files from four different Arabic countries, Iraq, Saudi Arabia, Syria, and Lebanon as shown in Table 3. Most of the open published files in the Arab region are either in PDF or Excel format. The reasons for choosing XLS format were data fidelity, ability to source from a wider range of public sector domains, and to have increased value that comes from many information linkages. The authors believe that for many years to come, more drug data will be published in XLS format in the Arab countries.

The above selected datasets are open data published by health ministries or equivalent bodies in the respected governments, these datasets are regularly updates (usually after a 2 year period). As it can be noticed from the difference in number of columns that the structure of the datasets is not unified same which enforces the unifications and mapping of data.

The data quality of the selected files is too low, e.g.; most of excel documents do not represent the generic name or their ATC code which makes the data almost unusable for further transformation. Lebanon and Saudi Arabia have built a generic online drug database on the following links:

- Saudi Arabia - https://www.sfda.gov.sa/en/drug/search/Pages/default.aspx,
- Lebanon - https://www.moph.gov.lb/en/Drugs/index/3/4848/lebanon-national-drug-index-lndi-

TABLE 3. SELECTED ARABIC OPEN DRUG DATASETS

Country	DataSet URI	No. Of Tuples	No. of Columns
Iraq	www.iraqipharm.com/upfiles/drug/dreg.xls	9090	9
Lebanon	https://moph.gov.lb/userfiles/files/HealthCareSystem/.../7../WebMarketed20170307.xls	5822	15
Saudi Arabia	https://www.sfda.gov.sa/en/drug/search/pages/default.aspx	6386	10
Syria	www.moh.gov.sy/LinkClick.aspx	9375	7

These two databases contain 13,445 records. In order to gather the data in HTML format, the authors built HTML

Crawlers based on JSOUP5 which is a Java library for extracting and manipulating data, it iterates drug list link per link, and gather the presented information for each drug separately. Unfortunately, Syria and Iraq do not provide such databases, so the authors have to use their Excel files and to implement additional transformation to extract active substances.

In what follows we will describe the steps required in the process of transforming, linking, and publishing the Arabic drug data.

1. *Data Cleaning*

OpenRefine (http://openrefine.orgVersion 2.6-rc1) used to clean the selected data in order to make it coherent and ready for further operations according to the methodology. A well-organized cleaning operation minimizes inconsistencies and ensures data standardization among a verity of data sources.

2. *Ontology Definition and Data Mapping Schema*

Some of the ontologies and vocabularies which a data publisher needs to have in mind biomedical ontologies. The schema comprises classes and properties are Schema.org[6], DBpedia Ontology UMBEL[7], DICOM[8], and the DrugBank Ontology used, as well as other from the Schema.org vocabulary: the schema:Drug[9] class, along with a large set of properties which instances of the class can have, such as generic *drug name*, *code*, *active substances*, *non-proprietary Name*, *strength value*, *cost per unit*, *manufacturer*, *related drug*, *description*, *URL*, *license*, *etc*.

Additionally, in order to align the drug data with generic drugs from DrugBank, properties *brandName*, *genericName*, *atcCode and dosageForm* from the DrugBank Ontology were used. The relation rdfs:seeAlso can be used to annotate the links which the drug product entities will have to generic drug entities from the LOD Cloud dataset.

The nodes are linked according to the relations these classes, tables or groups have between them. There exist a few tools for ontology and vocabulary discovery which should be used in this operation such as Linked Open Vocabularies (LOV, http://lov.okfn.org/) and DERI Vocabularies (http://datahub.io).

3. Data Conversion

Create RDF dataset: The previously mapped schema can produce RDF graph by using RDF-extension of LODRefine tool. This step transforms raw data into RDF dataset based on a serialization format.

Transformation process can be executed in many different ways, and with various software tools, e.g. OpenRefine (which the authors used), RDF Mapping Language[10], and XLWrap[11], among others.

Interlinking: LODRefine was used for reconciliation to make interlinking of the data. In this case, columns atcCode, genericName1, activeSubstance1, activeSubstance2 and activeSubstance3 reconciliated with DBpedia. This operation enables interoperability between organization data and the Web through establishing semantic links between the source dataset (organization data) with related datasets on the Web.

Link discovery activity can be performed in manual, semi-automated or fully-automated modes which help to discover links between the source and target datasets since the manual mode is tedious, error-prone, and time-consuming, and the fully-automated mode is currently unavailable, the semi-automated mode is preferred and reliable.

Link generation activity generates the links in RDF format using *rdfs:seeAlso* or *owl:sameAs* predicates. The activities of link discovery and link generation are performed sequentially for each data source.

The last activity within the interlinking stage is the generation of overall link statistics which showcase the total number of links generated between the source and target data sources.

Storage/SPARQL Endpoints: OpenLink virtuoso[12] server (version 06.01.3127) on Linux (x86_64-pc-Linux-gnu), Single Server Edition used to run the *SPARQL endpoint Queries:* http://aldda.b1.finki.ukim.mk/sparql

Publication: RDF graph can be accessed on the following link: http://aldda.b1.finki.ukim.mk/. For publishing Linked Data on the web, a Linked Data API is needed which makes a connection with the database to answer specific queries.

The HTTP endpoint is a webpage that forms the interface. To make a Web application, a REST API is used. This REST API gives the possibility to give the Linked Data in various formats back to the user, depending on the user's requirements.

The Linked Data can be made visible in HTML on a website as HTTP links, or as RDF data in a browser or the most user-friendly: a graphic visualization in a Web application.

4. Quality Assessment

The quality of datasets usually needs assessment of its quality in all previous stages; quality assessment is an ongoing operation in all stages as the quality of content of the document web varies. In the methodology, the authors strongly recommend assessing quality at every stage of the transformation process based on characteristics such as accuracy, consistency, and relevancy. As a future work, this stage is going to be elaborated further, in terms of quality assessment and a framework application.

[5] http://jsoup.org
[6] https://schema.org/
[7] http://umbel.org/
[8] https://www.dicomstandard.org/
[9] https://health-lifesci.schema.org/Drug
[10] http://rml.io/spec.html
[11] http://xlwrap.sourceforge.net/
[12] https://virtuoso.openlinksw.com/

5. Visualization and Querying.

After publishing the data and becomes available to the web application, the data can be displayed. Because data are available in a raw/un-interpreted form, there are many visualization opportunities in the web application. When turning data into information for the application user, freely available libraries can be used that offer diverse types of visualization such as a table or in a diagram formatted in different ways. Visualization libraries and other libraries can be used to enable a user to interact with data.

V. Discussion

Web drug data availability in some Arabic countries is basically public as a 2-star format data i.e. Excel or PDF format. Most of the available drug data is provided in the English language with a few columns in Arabic, this is due to the fact that the English language is widespread among physicians and pharmacists, and also English is the predominant language in communications between physicians and pharmacists.

In this search for drug data in Arabic countries, authors selected four drug data files from four different Arabic countries, Iraq, Syria, Saudi Arabia, and Lebanon.

Following the authors' proposal described above, datasets converted and interlinked with other available data including DBpedia have been uploaded to a Virtuoso store. Authors selected to use the OpenLink virtuoso[13] server (version 06.01.3127) on Linux (x86_64-pc-linux-gnu), Single Server Edition to run our *SPARQL endpoint Queries:* http://aldda.b1.finki.ukim.mk/sparql.

However, authors decided to use the RDF as their dataset format, because it is recommended by W3C, and has advantages, such as the provision of an extensible schema, self-describing data, de-referenceable URIs, and, as RDF links are typed, enable interoperability, structured, and safe linking of different datasets.

Before starting to convert out XLS to RDF, the authors selected target ontology to describe the drugs contained in the drug availability dataset. Authors selected to use the Linked Drugs ontology, Schema.org[14] vocabulary, and DBpedia as they cover the needed properties and provide easier interlinking possibilities for further transformation.

The Web Ontology Language allows complex logical reasoning and consistency checking of RDF/OWL resources. These reasoning capabilities helped the authors to harmonize the heterogeneous data structures found in the input datasets.

The authors transformed the selected drug data into five-star LOD and established relations in the RDF graph towards outside entities including the DBpedia and DrugBank.

The authors selected the 'owl:sameAs' relation to relate the drugs in the Arabic dataset with the entities in the Linked Drugs dataset and assume that the two drug descriptions refer to the same real-world entity.

In this paper, the authors presented the transformation process of 2-star drug data from selected Arabic countries published on various websites, into a 5-star Linked Open Data, connected to the DrugBank and DBpedia. The overall count of the distinct data was 31,906 drugs, and there 23,971 interlinked drugs to DBpedia.

The authors also provided use-cases which give examples of how the data from the Health Insurance Fund and DrugBank can be used, in order to provide application developers with mechanisms and ideas for retrieving distributed data in various formats.

The paper discussed the process of integrating different datasets and the possibility to build a *Arabic Linked Drug Data Application* on top of the Virtuoso RDF store. The paper emphasizes the advantages and issues of using the Linked Data approach in the pharmaceutical/drug industry.

The future work will include implementation of a stable and open-source version of a Java web application based on AngularJS[15], an open-source web applications that will allow the end-user to fully explore and assess the quality of the consolidated dataset [2], if possible, to repair the errors observed in the Arabic Linked Drug dataset.

Acknowledgment

The research presented in this paper is partly financed by the Ministry of Science and Technological Development of the Republic of Serbia (SOFIA project, Pr. No: TR-32010) and partly by the EU project LAMBDA (GA No. 809965).

References

[1] Auer S, *et al*., "Managing the Life-Cycle of Linked Data with the LOD2 Stack". In: The Semantic Web-ISWC 2012. Boston: Springer Berlin Heidelberg, pp. 1–16, 2012.

[2] Lackshen, G., Janev, V., Vraneš, S. (2018). Quality Assessment of Arabic DBpedia. In Proc. of 8th International Conference on Web Intelligence, Mining and Semantics. June 25 – 27 2018, Novi Sad, Serbia. ACM New York, NY, USA DOI: https://doi.org/10.1145/3227609.3227675

[3] Jentzsch A., *et al*., *Linking Open Drug Data*, *Triplification Challenge of the International Conference on Semantic Systems*, 2009.

[4] Jovanovik M. and Trajanov D., *Consolidating drug data on a global scale using linked data*. Journal of Biomedical Semantics, 8(3), 2017.

[5] W3C, *Best Practices for Publishing Linked Data*, 2016 http://www.w3.org/TR/ld-bp/

[6] Hyland B, Wood D., *The Joy of Data: A Cookbook for Publishing Linked Government Data on the Web*. In: Linking Government Data, New York: Springer New York;. 2011, pp. 3–26.

[7] Hausenblas M., *Linked Data Life Cycles*, 2016. http://www.slideshare.net/mediasemanticweb/linked-data-life-cycles.

[8] Cai, L. and Zhu, Y. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era, Data Science Journal. 14, p. 2, 2005.

[13] https://virtuoso.openlinksw.com/

[14] www.schema.org

[15] https://angularjs.org/