Computer Science and Information Systems 18(3):729–748 https://doi.org/10.2298/CSIS200510047L

Arabic Linked Drug Dataset Consolidating and Publishing

Guma Lakshen[1], Valentina Janev[2], and Sanja Vraneš[2]

[1] School of Electrical Engineering, University of Belgrade,
11000 Belgrade, Serbia
jlackshen65@yahoo.com
[2] Mihajlo Pupin Institute, University of Belgrade,
11060 Belgrade, Serbia
{valentina.janev, sanja.vranes}@pupin.rs

Abstract. The paper examines the process of creating and publishing an Arabic Linked Drug Dataset based on open drug datasets from selected Arabic countries and discusses quality issues considered in the linked data lifecycle when establishing a semantic Data Lake in the pharmaceutical domain. Through representation of the data in an open machine-readable format, the approach provides an optimum solution for information and dissemination of data and for building specialized applications. Authors contribute to opening the drug datasets from Arabic countries, interlinking the data with diverse repositories such as DrugBank, and DBpedia, and publishing it in a standard open manner that allows further integration and building different business services on top of the integrated data. This paper showcases how drug industry can take full advantage of the emerging trends for building competitive advantages. However, as is elaborated in this paper, better understanding of the specifics of the Arabic language is needed in order to extend the usage of linked data technologies in Arabic companies.

Keywords: drug management applications; Linked Data; methodology; open ecosystems; quality assessment.

1. Introduction

Today, data is growing at a tremendous rate on the Web, and is expected to reach 35 Zettabytes (1 ZB= 10^{21} bytes) by the end of 2019, and exceeds 175 zettabytes by 2025 [1]. This amount of data creates new opportunities for modern enterprises, especially in the context of analyzing value chains in a broader sense. The value chain considered in this study is the one presented in Fig. 1 that can be divided into 3 layers:

- Data sources layer, where different data sources and systems generate data. The interconnected systems in this layer are property of the organization or its partners, or the data is freely available on the Web.
- Data management layer, where the data is acquired via customized interfaces or crawled from the Web and transmitted using interconnected networks into storage

data centers. The data management layer in a modern data ecosystem is composed of data lakes and data warehouses.

– Data analytics and business intelligence layer, which refers to the application of artificial intelligence, mining algorithms, machine learning, and deep learning to process the data and extract useful knowledge for better decision making. Additionally, data visualization tools are used for visually examining processed data.

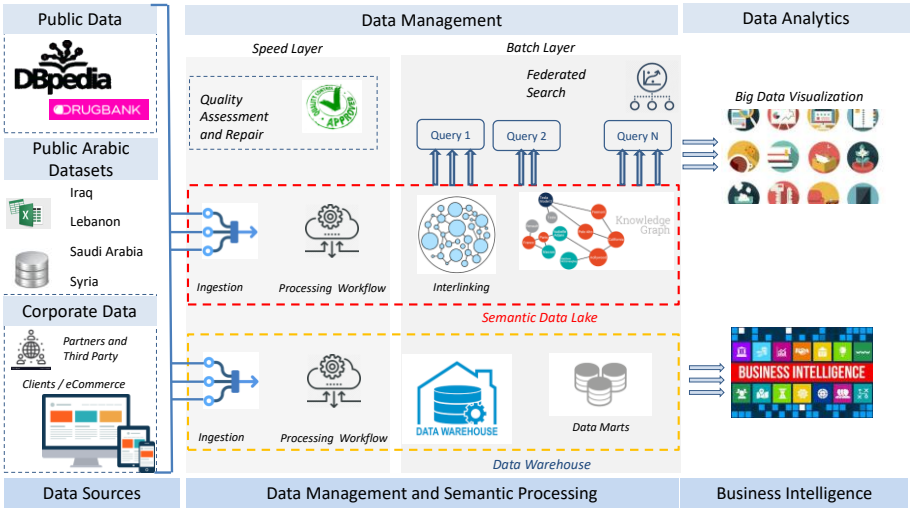

Fig. 1. Modern data ecosystem

The development of business intelligence services is simple when all data sources collect information based on unified file formats and the data is uploaded to a data warehouse. However, the development of a distributed software system requires the interaction of services and the use of resources from diverse organisations throughout the Web [2]. The biggest challenge that enterprises face is the undefined and unpredictable nature of data appearing in multiple formats. Additionally, in order to gain competitive advantage over their business rivals, the companies utilize open data resources that are free from restrictions and can be reused, redistributed, and can provide immediate information and insights. Thus, in a modern data ecosystem, data lakes and data warehouses are both widely used for storing big data. A *data warehouse* [3] is a repository for structured, filtered data that has already been processed for a specific purpose. A *data lake* is a large, raw data repository that stores and manages the company data bearing any format. The concept of data lake was introduced in the last decade in order to address issues related to processing big data [4]. Moreover, recently, the *semantic data lakes* [5] are introduced as an extension of the data lake supplying it with a semantic middleware, which allows the uniform access to original heterogeneous data sources. The data management life cycle (see Fig. 1) is divided into two parts. Data pre-processing activities like data integration, enrichment, transformation, reduction, and cleansing occur in the speed layer, while maintaining the knowledge graphs (part of the *semantic data lake*) and data marts is part of the batch layer. In addition to the speed and batch layers, in big data applications, a third layer is often added named (merged

serving layer that makes the combined data available for data analysis and reporting, see also Fig. 3.

In the last decade, more and more corporations have introduced semantic processing technologies also to improve the interoperability, i.e., they use the *Linked data principles and standards* recommended by the W3C consortium [6]. The use of common data models provides a standard way to store and query the data and, furthermore, creates an opportunity to build a virtual middleware under which the heterogeneous formats are homogenized on-the-fly without data transformation or materialization.

Taking the drug industry and drug management as an example, this study was motivated by the following research questions (related to operations in the depicted semantic layer, see red rectangular in Fig. 1):

- What are the benefits from integrating freely available data sources (e.g., DBpedia) into the existing business value chain and what are the drawbacks of this approach?
- What is the quality of open data e.g. the Arabic DBpedia?
- How can business intelligence services (e.g. a search operation) be implemented on top of a semantic drug data lake?

The paper is structured as follows. Section 2 presents the research framework and proposes a methodology for consolidating heterogeneous data sources using the Linked Data principles [6], [7]. Section 3 presents the process of transforming selected Arabic two-star drug datasets published on various websites, into a five-star linked open data (LOD)[1], connected to the DBpedia [8] and DrugBank [9]. The overall count of the distinct data is 31,906 drugs, while 23,971 drugs are interlinked to DBpedia. The proposed methodology advances the state-of-the-art taking into account quality issues and specifics for the Arabic language and provide examples of how the drug data lake (knowledge graph interlinked with DBpedia[2] and DrugBank[3]) can be analyzed with business objectives in mind (retrieval of drug information). Section 4 discusses the results and presents the main conclusions. Section 5 concludes the article.

2. Research Overview: Building Linked Data Application on Top of Arabic Open Data

In the drug industry, the rapidly increasing amount of data on the Web opens new opportunities for integrating and enhancing drug knowledge on a global scale. As far as medical data available in the Arab region, there are only a handful of Arabic drug applications such as Webteb[4], Altibbi[5], 123esaaf[6], Kuwait Pharmacy KP[7] which provides their services in Arabic and English, but unfortunately, the data is not open and most are

[1] https://www.w3.org/2011/gld/wiki/5_Star_Linked_Data
[2] https://wiki.dbpedia.org/
[3] https://www.drugbank.ca/
[4] https://www.webteb.com/aboutusen
[5] https://www.altibbi.com/
[6] https://www.123esaaf.com/
[7] http://www.kuwaitpharmacy.com/default.aspx

not free. Arabic language content on the Web is less than 3%. The situation is even worse regarding Arabic open data, linked data, and open linked data on drugs.

This limitation of Arabic content motivated the authors to propose a solution that will enable Arabic-speaking end users to benefit from private datasets interlinked with public data and local data enriched with information from the Web. The goal of the innovative Arabic drug application proposed in this study is to enable end-users to answer inquiries about drug availability in the open datasets and to enrich the local data store with information from the Web. Examples of key business queries are:

1. For a particular drug, retrieve relative information in the Arabic language (if exists) from other identified datasets, such as DrugBank and DBpedia.
2. For a particular drug, retrieve equivalent drugs, and compare their active ingredients, contradictions, and prices.
3. For a particular drug, retrieve valuable information about equivalent drugs with different commercial names, manufacturers, strengths, forms, prices, etc.
4. For a particular drug, retrieve its reference information to highlight possible contradiction, e.g., in combination with other drugs, allergies, or special cases (e.g., pregnancy).
5. For a particular active ingredient, retrieve advanced clinical information, i.e., pharmacological action, pharmacokinetics, etc.
6. For a particular drug, retrieve its cost, manufacturer, and country.

The authors also propose to split the implementation of a linked data application development into three software development phases as is presented in Table 1.

Table 1. Linked data application phases

Phase	Description
Initialization	Business objectives and requirements: Requirement specification, technical characterization, and setting up of the demo site; Establishing acceptance (success) criteria for pilot applications validation based on performance characteristics, usability, as well as EU and national regulations (e.g., related to data access and security measures); Data categorization and description: Analysis of the datasets to be published in linked data format and selection of vocabularies and development other specifications for metadata description. *Example.* In addition to corporate data, the targeted data is selected from the Arabic drug datasets (Iraq, Saudi Arabia, Syria, and Lebanon) along with the public datasets (DrugBank and DBpedia). Appropriate vocabularies are selected or developed, and mapping rules are defined.
Innovation	Integrating datasets in the form of a knowledge graph: Data access, transformation, and enrichment. For instance, in this phase, the data lake is established, and semantic processing is performed, which includes all the stages of data preparing, modeling, and conversion. At each stage, quality issues are revised, and if the quality is not satisfactory, the appropriate stage is revisited. After the transformation, master data is stored for subsequent use. Generic component selection and tool customization for the pilot applications: Customization of linked data components for use in the targeted domain. *Example.* In this phase, tools for federated search and data are selected. Additionally, big data analytics tools are selected, and custom visualization and user interfaces are created [8].
Specific tools development and Validation	In this phase open-source tools are validated for reuse; feedback is provided for improving the solution components; and new interfaces are built.

3. Validation of Software Development Methodology

3.1. Initialization (Stage I): Data Preparation

Data Selection. 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 2. Most of the open published files in the Arab region are either in PDF or XLS 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 selected datasets are open data published by health ministries or equivalent bodies in the respected governments. They are regularly updated, usually after a two-year period. As it can be noticed from the difference in the number of columns, the structure of the datasets is not unified, which makes the unification and mapping of data necessary.

Table 2. Selected Arabic open drug datasets

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

Data Analysis. The data quality (DQ) of the selected files is too low, e.g., most XLS documents do not represent the generic name or their ATC code, which makes the data almost unusable for further transformation. However, the data from Lebanon and Saudi Arabia is in a form of generic online drug database, see Table 2.

These two databases contain 13,445 records. In order to gather the data in HTML format, the authors built HTML Crawlers based on JSOUP, which is a Java library for extracting and manipulating data. It iterates through the drug list (link by link), gathering information for each drug separately. Unfortunately, Syria and Iraq do not provide such databases, so the authors have to use their XLS files and implement additional transformations to extract active ingredient information.

Data Cleaning. OpenRefine[8] (version 2.6-rc1) was 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.

[8] http://openrefine.org

Quality Assessment. In what follows, we will describe the process for transforming, linking, and publishing the Arabic drug data. When it comes to quality assessment of the DBpedia Arabic Chapter, there are problems specific to the Arabic language that result in:

1. Presentation of characters as symbols via web browsers due to errors during the extraction process.
2. Wrong values in numerical data, due to the use of Hindu numerals in some Arabic sources.
3. Occurrence of different names for the same attribute, for instance, the birth date attribute appears in various info-boxes by different names: one time as "تاريخ الميلاد" [birth date], another time as"تاريخ الولادة" [delivery date], the third time as "الميلاد" [birth].
4. Inconsistency of names between the infobox and its template; for instance, there is a template called "مدينة" [city] while the infobox name is called "معلومات مدينة" [city information].
5. Geo-names templates formatting problems when placed in the infobox.
6. Errors in *owl:sameAs* relations and problems in identifying the *owl:sameAs* relations due to heterogeneity in different data sources.

However, some of the problems present in other DBpedia chapters are also identified in the Arabic chapter, specifically:

1. Wrong Wikipedia Infobox information; for example, the height of minaret of the grand mosque in Mecca (the most valuable mosque for all Muslims) is given as 1.89 m, where the correct height is 89 m.
2. Mapping problems from Wikipedia, such as unavailability of infoboxes for many Arabic articles; for example, "Man-made River in Libya النهر الصناعي في ليبيا," it is considered as the biggest water pipeline project in the world, or not containing all the desired information.
3. Object values incompletely or incorrectly extracted.
4. Data type incorrectly extracted.
5. Some templates may be more abstract, thus cannot map to a specific class.
6. Some templates are unused or missing inside the articles.

3.2. Innovation (Stage 2 and Stage III): Integrating Datasets in the Form of a Knowledge Graph

Ontology Definition and Data Mapping Schema. The ontology development was based on re-use of classes and properties from existing ontologies and vocabularies including Schema.org vocabulary[9], DBpedia Ontology[10], UMBEL (Upper Mapping and Binding Exchange Layer)[11], DICOM (Digital Imaging and Communications in Medicine)[12], and DrugBank. Each instances of the drug class has properties such as generic drug name, code, active substances, non-proprietary name, strength value, cost

[9] https://schema.org/
[10] https://wiki.dbpedia.org/services-resources/ontology
[11] http://umbel.org/
[12] https://www.dicomstandard.org/

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)[13] and DERI Vocabularies[14].

Data Conversion. *Create RDF dataset:* The previously mapped schema can produce an RDF graph by using RDF-extension of LODRefine tool. This step transforms raw data into RDF dataset based on a serialization format. The transformation process can be executed in many different ways and with various software tools, e.g., OpenRefine (which the authors used), RDF Mapping Language[15], and XLWrap[16] which is a Spreadsheet-to-RDF Wrapper, among others.

*Interlinking***.** LODRefine[17] was used for reconciliation in interlinking the data. In this case, columns *atcCode*, *genericName1*, *activeSubstance1*, *activeSubstance2* and *activeSubstance3* reconciled 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 can be performed in manual, semi-automated, or fully-automated modes to help 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 yields 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 and Publishing. OpenLink Virtuoso server (version 06.01.3127)[18] on Linux (x86_64-pc-Linux-gnu), Single Server Edition have been used to run the SPARQL endpoint queries: http://aldda.b1.finki.ukim.mk/sparql. 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. A REST API is used to make a web application. It makes it possible to give the linked data back to the user in various formats, 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 a graphic visualization in a web application, which would be the most user-friendly.

[13] http://lov.okfn.org/
[14] http://datahub.io
[15] https://github.com/RMLio
[16] http://xlwrap.sourceforge.net/
[17] https://sourceforge.net/projects/lodrefine/
[18] https://github.com/openlink/virtuoso-opensource

3.3. Specific Tools Development and Validation (Stage II and Stage III)

Tools for Quality Assessment. In our approach [10], [11], quality assessment is an ongoing operation in all stages as the quality of the content of the document on the Web varies, see also Fig. 2. The authors strongly recommend assessing quality at every stage of the transformation process based on characteristics such as accuracy, consistency, and relevancy. Therefore, we have developed an evaluation scheme that addresses the DQ before starting data analytics. It is carried out by estimating the quality of data attributes or features by applying a dimension metric to measure the quality characterized by its accuracy, completeness, and consistency.

The expected result is DQ assessment suggestions indicating the quality constraints that will increase or decrease the DQ. The authors believe also that DQ must be handled in many other phases of the big data lifecycle. In our approach, we distinguish between quality on data level and quality on metadata level. The data pre-processing improves DQ by executing many tasks and activities such as data transformation, integration, fusion, and normalization.

Example. For every quality dimension, quantification and measurement are needed (see the discussion on dimensions in Section 3.1). Therefore, metrics have been defined and linked to particular dimensions. Usually, most metrics used for measuring DQ are within a range from 0 to 1, with 0 representing incorrect value and 1 representing a correct value. Dimensions such as accuracy, completeness, and consistency, among others, are calculated by the function $M\_D = 1 - (N_{iv}/N_{tv})$, where *M_D* is the metric for a given dimension, N_{iv} is the count of incorrect values, and N_{tv} is the total number of values for the dimension concerned. Regarding DQ dimensions relevant for quality assessment of Arabic DBpedia, we have identified three dimensions accuracy, consistency, and relevancy, as shown in Table 3.

Table 3. Data Quality dimensions relevant for quality assessment of Arabic DBpedia (*Specific to DBpedia, **Specific to Arabic DBpedia)

Category	Sub-category
Accuracy	• Incorrectly extracted triple • Special template not properly recognized* • Wrong values in numerical data (due to dual numbering used) **
	Incorrectly extracted data type
	Implicit relationship between attributes • One/ Several fact encoded in one/several attributes* • Attribute value computed from another attribute value**
Consistency	• Inconsistency in representation of number values**
Relevancy	Irrelevant information extracted • Extraction of attributes containing layout information** • Redundant attribute values • Image related information* • Other irrelevant information

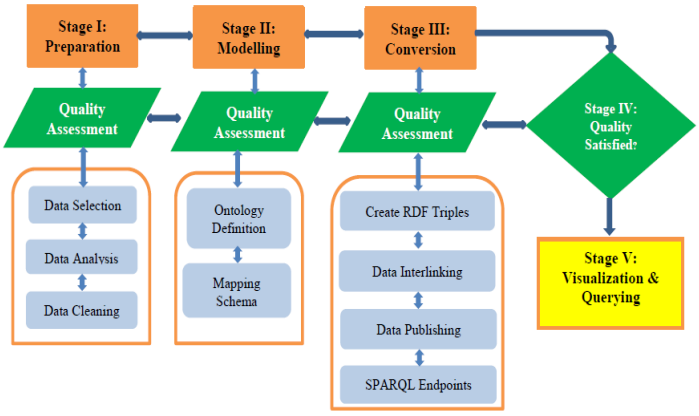

Fig. 2. A novel linked data methodology with a focus on quality assessment

Tool for Workflow Automation. The processing steps discussed so far refer to the initial load of the knowledge graph available online for experimental purposes at this location[19]. Currently, underway is testing of the solution and deployment of the adopted tools (LODRefine, OpenLink Virtuoso, PoolParty UnifiedViews for a client from Lybia. The PoolParty UnifiedViews (relevant for the speed layer in the Big Data architecture presented in Fig. 1) is considered for automation of the Extract-Transform-Load processes. The UnifiedViews's pipeline shall integrate also the custom quality assurance services discussed above.

3.4. Visualization and Querying (Stage IV)

After publishing the data on the Web in a form of a knowledge graph, it becomes available to other web applications for retrieval and visualization [12]. Using standard vocabularies for modeling allows end users to use different visualization approaches, e.g., freely available libraries can be used that offer diverse types of visualization, such as a table or in a diagram formatted in different ways as shown in Fig. 3. Custom visualization and query applications enable the user to interact with the data. In order to visualize the statistics about drug types and/or manufacturers, we use the exploratory spatial-temporal analysis (ESTA-LD) [12] tool[20]. The tool enables us to select the endpoint from where the data should be retrieved. This Section gives few examples of SPARQL queries that answer the business questions introduced in Section 1.

[19] http://aldda.b1.finki.ukim.mk/sparql, http://aldda.b1.finki.ukim.mk.
[20] http://geoknow.imp.bg.ac.rs/ESTA-LD

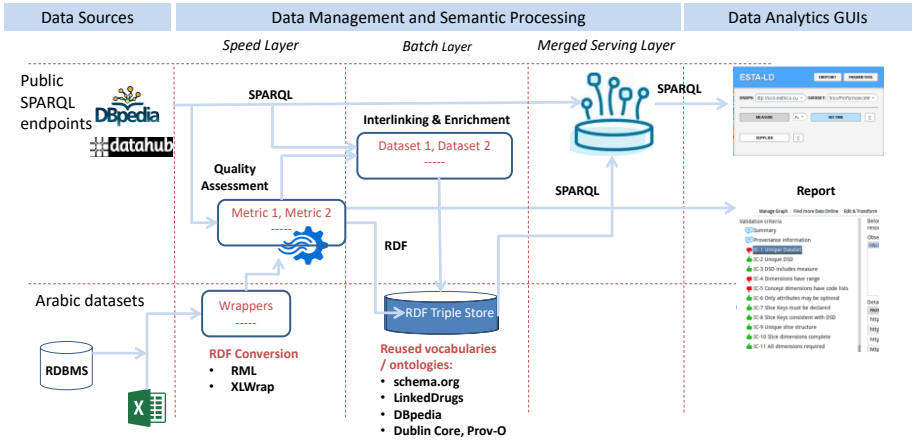

Fig. 3. Knowledge graph visualization and querying

Example. Query: Count all distinct drugs

```
PREFIX dbo: <http://dbpedia.org/ontology/>
SELECT count distinct ?drug
FROM <http://aldda.b1.finki.ukim.mk/lod/data/drugs>
WHERE
   { ?drug a <http://schema.org/Drug>  }
```

Output: 31906 distinct drugs

- Query: Count all interlinked drugs

```
PREFIX dbo: <http://dbpedia.org/ontology/>
SELECT count distinct ?drug
FROM <http://aldda.b1.finki.ukim.mk/lod/data/drugs>
WHERE
   { ?drug a <http://schema.org/Drug> .
    ?drug rdfs:seeAlso ?seeAlso }
```

Output: 23971 interlinked drugs

It is notable that >75% of the merged datasets are interlinked with DBpedia and can obtain additional information regarding drugs from DBpedia.

- Query: Extract abstract info from DBpedia in Arabic language for the ‘taxol’ which is an Organic composite similar to the ‘paclitaxel’ drug

```
prefix dbo: <http://dbpedia.org/ontology/>
prefix drugbank: <http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugbank/>
SELECT * WHERE {
?drug a <http://schema.org/Drug>   .
?drug drugbank:genericName ?genericName .
?drug rdfs:seeAlso ?seeAlso .
{ SERVICE<http://dbpedia.org/sparql>
{ ?seeAlso dbo:abstract ?abstract  } }
FILTER (?genericName = ‘paclitaxel’)
FILTER (langMatches(lang(?abstract), "ar")) }
```

- Output

"محضر من لحاء ، وهو مركب taxol في 1988 توصل الباحثون في جامعة جونز هوبكنز إلى أن تاكسول بسرطان حاد في المبيض. كما اقترح الباحثون شجر الطقسوس بالمحيط الهادي ، يمكن أن يفيد النساء المصابات للسرطان في هيوسطن أن مادة تاكسول يمكن أن تفيد السيدات المصابات سنة 1991 في مركز أندرسون الثدي أيضاً. في دراسات تمت على 25 سيدة مصابة بسرطان متقدم في الثدي ولمتتمكن من الاستجابة بسرطان بعد تسع شهور من العلاج التجريبي للعلاج الكيمائي ، شعر غالبية السيدات بانكماش الورم".@ar

- Query: Equivalent drugs comparison

```
prefix dbo: <http://dbpedia.org/ontology/>
prefix drugbank: <http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugbank/>
prefix schema:<http://schema.org/>
prefix dbp: <http://dbpedia.org/ontology/>
SELECT distinct ?drug1, ?drug1GenericName, ?drug1ManufacturerLegalName, ?drug1ActiveIngredient, CONCAT(str(?drug1CostPerUnit),' ',?drug1CostCurrency) as ?drug1CostFull, ?drug1AddressCountry,
?drug2,?drug2GenericName, ?drug2ManufacturerLegalName, ?drug2ActiveIngredient, CONCAT(str(?drug2CostPerUnit),' ',?drug2CostCurrency) as ?drug2CostFull, ?drug2AddressCountry WHERE {
?drug a <http://schema.org/Drug> .
?drug drugbank:genericName ?drug1GenericName .
?drug schema:addressCountry ?drug1AddressCountry .
?drug schema:cost ?drug1Cost .
?drug schema:manufacturer ?drug1Manufacturer .
?drug1Manufacturer schema:legalName ?drug1ManufacturerLegalName .
?drug schema:activeIngredient ?drug1ActiveIngredient .
?drug1Cost schema:costPerUnit ?drug1CostPerUnit .
?drug1Cost schema:costCurrency ?drug1CostCurrency .
?drug rdfs:seeAlso ?seeAlso .
?drug2 rdfs:seeAlso ?seeAlso .
?drug2 drugbank:genericName ?drug2GenericName .
?drug2 schema:addressCountry ?drug2AddressCountry .
?drug2 schema:cost ?drug2Cost .
?drug2 schema:manufacturer ?drug2Manufacturer .
?drug2Manufacturer schema:legalName ?drug2ManufacturerLegalName .
?drug2 schema:activeIngredient ?drug2ActiveIngredient .
?drug2Cost schema:costPerUnit ?drug2CostPerUnit .
?drug2Cost schema:costCurrency ?drug2CostCurrency .
FILTER (?drug != ?drug2)}
```

- Output:

	Drug1	Drug2
Drug Number	aldda.b1.finki.ukim.mk/lod/data/drugs#35704	aldda.b1.finki.ukim.mk/lod/data/drugs#36482
GenericName	glimepiride	metformin and sulfonamides
ManufacturerLegalName	Sadco	Benta Trading Co s.a.l.
ActiveIngredient	Glimepiride	Metformin HCl
CostFull	12415.0 L.L	28800.0 L.L
AddressCountry	LB	LB

- Query: Drugs with different brand name comparison.

```
prefix dbo: <http://dbpedia.org/ontology/>
prefix drugbank: <http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugbank/>
prefix schema:<http://schema.org/>
prefix dbp: <http://dbpedia.org/ontology/>
SELECT ?drug1BrandName,?drug1GenericName, ?drug1ManufacturerLegalName,
?drug1ActiveIngredient, ?drug1DosageForm, CONCAT(str(?drug1CostPerUnit),'
',?drug1CostCurrency) as ?drug1CostFull, ?drug1AddressCountry,
?drug2BrandName,?drug2GenericName, ?drug2ManufacturerLegalName,
?drug2ActiveIngredient, ?drug2DosageForm, CONCAT(str(?drug2CostPerUnit),'
',?drug2CostCurrency) as ?drug2CostFull, ?drug2AddressCountry WHERE {
?drug a <http://schema.org/Drug> .
?drug drugbank:brandName ?drug1BrandName .
?drug drugbank:genericName ?drug1GenericName .
?drug schema:addressCountry ?drug1AddressCountry .
?drug schema:cost ?drug1Cost .
?drug schema:manufacturer ?drug1Manufacturer .
?drug1Manufacturer schema:legalName ?drug1ManufacturerLegalName .
OPTIONAL {
?drug drugbank:dosageForm ?drug1DosageForm }
?drug schema:activeIngredient ?drug1ActiveIngredient .
?drug1Cost schema:costPerUnit ?drug1CostPerUnit .
?drug1Cost schema:costCurrency ?drug1CostCurrency .
?drug rdfs:seeAlso ?seeAlso .
?drug2 rdfs:seeAlso ?seeAlso .
?drug2 drugbank:brandName ?drug2BrandName .
?drug2 drugbank:genericName ?drug2GenericName .
?drug2 schema:addressCountry ?drug2AddressCountry .
?drug2 schema:cost ?drug2Cost .
?drug2 schema:manufacturer ?drug2Manufacturer .
?drug2Manufacturer schema:legalName ?drug2ManufacturerLegalName .
?drug2 schema:activeIngredient ?drug2ActiveIngredient .
OPTIONAL { ?drug2 schema:availableStrength ?drug2Strength .}
OPTIONAL {?drug2 drugbank:dosageForm ?drug2DosageForm }
?drug2Cost schema:costPerUnit ?drug2CostPerUnit .
?drug2Cost schema:costCurrency ?drug2CostCurrency .
FILTER (?drug1BrandName != ?drug2BrandName &&
?drug1DosageForm != ?drug2DosageForm &&
?drug1ManufacturerLegalName
!=drug2ManufacturerLegalName)}
```

- Output

	Drug1	Drug2
BrandName	EBETREXAT	METOJECT
GenericName	methotrexate	methotrexate
ManufacturerLegalName	Codipha	Alfamed S.A.L.
ActiveIngredient	methotrexate	methotrexate
DosageForm	7.5mg/0.75ml	15mg/0.3ml
CostFull	32984.0 L.L	51182.0 L.L
AddressCountry	LB	LB

4. Discussion of Results: Analysis of Linked Data Methodologies

In literature, not many papers have dealt with linked data methodologies i.e., the process of generating, linking, publishing, and using linked data; to name a few: *W3C Best Practices for Publishing Linked Data* (W3C-Government Linked Data Working Group, 2014)[21] [13]; *A Cookbook for Publishing Linked Government Data on the Web* [14]; *Linked Data Life Cycles* [15]; *Guidelines for Publishing Government Linked Data* [16]; *Managing the Life-Cycle of Linked Data with the LOD2 Stack* [6]; and *Methodological Guidelines for Consolidating Drug Data* [17]; see Table 4 for a comparison. One of the first linked data methodologies was developed in the European research project LOD2 (Creating Knowledge out of interlinked Data, 2011-2014)[22] that was mainly dedicated to the publishing process, i.e., opening data in a machine-readable format and establishing the tools and technologies for interlinking and integrating heterogeneous data sources in general.

Jovanovik and Trajanov [17] 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." (page 4). Therefore, they proposed a new linked data methodology with a focus on reuse. It provides guidelines for data publishers defining reusable components in the form of tools, schemas, and services for the given domain (i.e., drug management).

The methodology presented in this paper meets the needs of the case study. Hence we suggest an approach to standardize the quality assessment of Linked Data lifecycle as is presented in Table 5.

[21] W3C Best Practices for Publishing Linked Data. http://www.w3.org/TR/ld-bp/ (2018)
[22] https://linkeddata.rs/project/LOD22010–2014

Table 4. Comparison of previous methodologies

Authors	Title / Steps	
W3C Government Linked Data Working Group (2014)	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)	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)	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)	Guidelines for Publishing Government Linked Data:	
	(1) Specify, (2) Model,	Initialization
	(3) Generate, (4) Publish,	Innovation
	(5) Exploit	Validation & Maintenance
Auer, et all. (2012)	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)	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

Table 5. The proposed methodologies

Guma Lakshen	Methodological guidelines for quality assessment of Linked Data:	
	(I) (1)Data Selection, (2) Data Analysis and (3) Data Cleaning,	Initialization
	(II) (4) Ontology Definition, (5) Mapping Scheme taking into consideration Quality metrics, (III) (6) Conversion into 5-star Linked Data taking into consideration the specific requirements of the Arabic language, and (7) Interlinking, (8) Publishing the Linked Data Dataset on the Web,	Innovation
	(IV) (9) Quality Assessment, (V) (10) Use-cases, Applications and Services.	Validation & Maintenance

4.1. Analysis of the Knowledge Graph Model and Transformation Tools

In this study, the authors decided to use the RDF, because it is recommended by W3C, and it has advantages, such as an extensible schema, self-describing data, de-referenceable URIs. Further on, since RDF links are typed, it enables good structure, interoperability, and safely linking different datasets.

Before converting XLS data to RDF, the authors selected the target ontology to describe the drugs contained in the drug availability dataset. Authors selected the LinkedDrugs[23] ontology [17], Schema.org vocabulary, and DBpedia, as they have the needed properties and provide easier interlinking possibilities for further transformation. The Web Ontology Language allows complex logical reasoning and consistency checking for RDF and OWL resources. These reasoning capabilities helped the authors to harmonize the heterogeneous data structures found in the input datasets.

Web drug data availability in some Arabic countries is basically public as a two-star format data, i.e., PDF or XLS format. Most of the available drug data is provided in the English language with a few columns in Arabic, this is because English is widely used among physicians and pharmacists; it is the predominant language in their communications.

Following the authors' proposal described above, the authors transformed the selected drug data into five-star linked open data and established relations in the RDF knowledge graph (31,906 drugs, more than 300 000 triples) toward outside entities, including the DBpedia and DrugBank. The *owl:sameAs* relation allows interlinking related drug descriptions that refer to the same real-world entity. For research purposes, the knowledge graph has been published via the SPARQL endpoint available at http://aldda.b1.finki.ukim.mk/sparql.

4.2. Quality Analysis of Integrated Open Data

Many authors have pointed out issues such as the completeness, conciseness, and consistency of open data. In 2014, Kontostas et al. [18] provided several automatic quality tests on LOD datasets based on patterns modeling various error cases, and they detected 63 million errors among 817 million triples. At the same time, Zaveri et al. [19, 20], conducted a user-driven quality evaluation which stated that DBpedia indeed has quality problems (e.g., around 12% of the evaluated triples had issues). They can be summarized as incorrect or missing values, incorrect data types, and incorrect links. Based on the survey, the authors developed a comprehensive quality assessment framework based on 18 quality dimensions and 69 metrics. Based on the work of Zaveri et al. [120], and the ISO 25012 DQ model, Radulović et al. [21] developed a linked data quality model and tested the model with DBpedia with a special focus on accessibility quality characteristics.

Based on the analysis of quality issues with DBpedia and the problems identified, the authors conclude that most important dimensions to be taken into consideration are the following.

[23] http://linkeddata.finki.ukim.mk/sparql

– Accuracy: triple incorrectly extracted, data type problems, errors in the implicit relationship between attributes.
– Consistency: representation of numerical values.
– Relevancy: irrelevant information extracted.

Different metrics were further defined, and web services were implemented to be used for data curation.

4.3. Proposal for Further Development of Quality Assessment Tools

There were several attempts in the past to design and implement a generic tool for linked data quality assessment. One of the first open-source frameworks for flexibly expressing quality assessment methods, as well as fusion methods, was Sieve (http://sieve.wbsg.de) Mendes et al. [22], released in 2012. As part of the Linked Data Integration Framework (LDIF; http://sieve.wbsg.de/), Sieve supports users in accessing data from the LOD cloud. Taking into consideration that DBpedia is a core element in the LOD cloud, in 2014, Kontokostas et al. enabled the RDFUnit Testing Suite (https://github.com/AKSW/RDFUnit) to run automatically-generated (i.e., based on a schema) and manually-generated test cases against an endpoint, e.g., the DBpedia SPARQL endpoint. Recognizing the large variety of DQ dimensions and measures, Luzzu (https://github.com/EIS-Bonn/Luzzu) [23], was developed at the same time to allow knowledgeable engineers without Java expertise to create quality metrics in a declarative manner. LOD Laundromat (http://lodlaundromat.org) was designed to help crawl the LOD cloud, converting all its contents in a standards-compliant way (i.e., gzipped N-Triples), as well as removing all data stains, such as syntax errors, duplicates, and blank nodes. TripleCheckMate (https://github.com/AKSW/TripleCheckMate) is a tool for crowdsourcing the assessment of Linked Open Data. It was developed for evaluating the correctness of DBpedia. TripleCheckMate provides an easy user interface with multiple resource assignment methods and a ready-to-use error classification scheme. The quality assessment methods implemented in these tools can be grouped into automatic, semi-automatic, manual, or crowd-sourced approaches. Initial results of the analysis and a comparison of the selected tools are provided in Table 6. However, as these tools have not been tested with the Arabic datasets yet, the quality assessment operations needed in our case study were implemented with custom web services.

Table 6. Comparison of open source quality assessment tools according to several attributes

Tool	Extensibility	Last Update	Collaboration	Cleaning
RDFUnit	SPARQL	03/2018	×	×
Luzzu	JAVA, LQML	07/2017	×	×
TripleCheck Mate	×	03/017	✔	×
Laundromat	SPARQL	05/2018	✔	✔
Sieve	XML	2014	×	✔

5. Concluding Remarks

Most of the available drug datasets nowadays are still provided in 2-star format and in English language due to the fact that the English language is widespread among physicians and pharmacists and also a predominant language in communications between physicians and pharmacists. In order to showcase the possibilities for large-scale integration of drug data, the authors proposed a piloting methodology and tested the approach with datasets from Arabic countries. The authors presented the transformation process of 2-star drug data into a 5-star Linked Open Data with DrugBank and DBpedia. The data is open for research purposes, while the OpenLink Virtuoso server (version 06.01.3127) on Linux (x86_64-pc-linux-gnu), Single Server Edition has been used to run the SPARQL endpoint (see http://aldda.b1.finki.ukim.mk/sparql).

The paper showcases the benefits from the Linked Data approach, in particular the possibility of enriching the private datasets with selected open data such as DBpedia. Main conclusion is that the Linked Data approach (1) contributes to the standardization on the metadata level and the semantic interoperability; (2) opens possibilities for improving the existing business value chain and insights by integration of valuable free information. However the quality issues in the Big Data ecosystems, Linked Drug Data in particular are still wide open for further study and evaluation, especially in the Arab countries.

The main research goal was to identify, collect, analyze, and evaluate the quality of selected drugs data sets, to allow quantifying and improving their value for the benefit of the user's especially with deficiencies in English language. The main contributions can be summarized as follows:

- This work introduced a modified process model based on previous methodologies shown in table 3 above.
- It is recommended to implement custom quality assessment services for transformation and processing in order to ensure that the process is conducted in high quality manner.
- For the first time, the paper discusses the issues with drug data from Arabic countries based on the selected four drug data files from four different Arabic countries, Iraq, Syria, Saudi Arabia, and Lebanon).

Taking into consideration the issues identified with quality of the open data (in particular, the issues with drug data from Arabic countries), the future work will include

implementation of additional services for repairing the errors observed in the Arabic Linked Drug dataset.

Acknowledgment. The research leading to these results has received funding from the European Community's H2020 Programme under grant agreement no 809965 (LAMBDA – Learning, Applying, Multiplying Big Data Analytics) and from the Ministry of Science and Technological Development of Republic of Serbia.

References

1. Patrizio, A.: IDC: Expect 175 zettabytes of data worldwide by 2025, Network World, Network World. https://www.networkworld.com/article/3325397/idc-expect-175-zettabytes-of-data-worldwide-by-2025.html (2018)
2. Aljazzaf, Z. M.: Modelling and measuring the quality of online services, Kuwait J. Sci. 42 (3), 134-157. (2015)
3. Kern, R., Kozierkiewicz, A., Pietranik, M.: The data richness estimation framework for federated data warehouse integration. Information Sciences, Volume 513, 2020, pp. 397-411. ISSN: 0020-0255. DOI: https://doi.org/10.1016/j.ins.2019.10.046 (2020)
4. Sawadogo, P., Darmont, J.: On data lake architectures and metadata management. Journal of Intelligent Information Systems. DOI: https://doi.org/10.1007/s10844-020-00608-7
5. Mami, M.N., Graux, D., Scerri, S., Jabeen, H., Auer, S., Lehmann, S.: Uniform Access to Multiform Data Lakes using SemanticTechnologies. Proceedings of the 21st International Conference on Information Integration and Web-based Applications & Services, December 2019, pp. 313–322. DOI: https://doi.org/10.1145/3366030.3366054 (2019)
6. Auer, S., A., Lorenz, B., Dirschl, C., Erling, O., Hausenblas, M., Isele, R., Lehmann, J., Martin, M., Mendes, P.N., Nuffelen, P.v., Stadler, C., Tramp, S. & Williams, H.: Managing the Life-Cycle of Linked Data with the LOD2 Stack. The Semantic Web-ISWC 2012. Boston: Springer Berlin Heidelberg: 1–16. (2012)
7. Berners-Lee, T.: Design issues: Linked data. http://www.w3.org/DesignIssues/LinkedData.html (2006)
8. Auer S., Bizer C., Kobilarov G., Lehmann J., Cyganiak R., Ives Z. (2007) DBpedia: A Nucleus for a Web of Open Data. In: Aberer K. et al. (eds) The Semantic Web. ISWC 2007, ASWC 2007. Lecture Notes in Computer Science, vol 4825. Springer, Berlin, Heidelberg.
9. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, Chang Z, Woolsey J. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res. 2006 Jan 1;34(Database issue):D668-72. DrugBank Release Version 5.1.7, https://www.drugbank.ca/releases/latest#open-data
10. Lackshen, G., Janev, V., Vraneš, S.: Quality Assessment of Arabic DBpedia. In R. Akerkar et all (Eds.) 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 (2018)
11. Lackshen, G., Janev, V., Vraneš, S.: Linking Open Drug Data: Lessons Learned. In K. Saeed et al. (Eds.): CISIM 2019, LNCS 11703, pp. 1–12, 2019. Springer International Publishing. https://doi.org/10.1007/978-3-030-28957-7 (2019)
12. Mijović, V. Janev, V., Paunović, D., Vraneš, S.: Exploratory Spatio-Temporal Analysis of Linked Statistical Data, Journal of Web Semantics, Web Semantics: Science, Services and Agents on the World Wide Web 41C (2016), pp. 1-8. ISSN: 15708268. DOI: https://doi.org/10.1016/j.websem.2016.10.002 (2016)

13. Janev, V., Mijović, V., Vraneš, V.: Using the Linked Data Approach in European e-Government Systems. International Journal on Semantic Web and Information Systems 14(2):27-46, April 2018. DOI: https://doi.org/10.4018/IJSWIS.2018040102 (2018).
14. 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: 3–26. (2016)
15. Hausenblas, M.: Linked Data Life Cycles. http://www.slideshare.net/mediasemanticweb/linked-data-life-cycles (2016)
16. Villaz´on-Terrazas, B. Vilches-Bl´azquez, L. Corcho, O. & G´omez-P´erez, A.: Methodological Guidelines for Publishing Government Linked Data Linking Government Data. In Linking Government Data. Springer NewYork, New York, NY. chapter 2: 27–49. (2011)
17. Jovanovik, M. & Trajanov, D.: Consolidating drug data on a global scale using linked data. Journal of Biomedical Semantics, 8(3). (2017)
18. Kontokostas, D. Westphal, P. Auer, S. Hellmann, S. Lehmann, J. & Cornelissen, R.: Test driven Evaluation of Linked Data Quality. In Proceeding of the 23rd International Conference on World Wide Web, pp. 747-758. New York, NY, USA. DOI: http://dx.doi.org/10.1145/2566486.2568002 (2014)
19. Zaveri, A. Kontokostas, D. Sherif, M.A. Bühmann, L. Morsey, M. Auer, S. & Lehmann, J. : User-driven Quality Evaluation of DBpedia. In Proceedings of the 9th International Conference on Semantic Systems. New York. USA:97–104. (2013)
20. Zaveri, A. Rula, A. Maurino, R. Pietrobon, R. Lehmann, J. & Auer, S.: Quality assessment for linked data: A survey. Semantic Web– Interoperability, Usability, Applicability. (2016)
21. Radulovic, F. Mihindukulasooriya, N. García-Castro, R. & Gómez-Pérez, A.: A Comprehensive Quality Model for Linked Data. Semantic Web – Interoperability, Usability, Applicability, Vol. 9, No. 1. Special issue on Quality Management of Semantic Web Assets (Data, Services and Systems), pp: 3-24. DOI: https://doi.org/10.3233/SW-170267 (2018)
22. Mendes, P.N. Mühleisen, H. & Bizer, C.: Sieve: Linked Data Quality Assessment and Fusion. In Proceedings of the 2012 Joint EDBT/ICDT Workshops, pp: 116-123. ACM, New York, NY, USA. DOI: http://dx.doi.org/10.1145/2320765.2320803 (2012)
23. Debattista, J. Auer, S & Lange, C.: Luzzu -- A Framework for Linked Data Quality Assessment. In Proceeding of the 2016 IEEE Tenth International Conference on Semantic Computing (ICSC), pp: 124-131. Laguna Hills, CA, 2016. IEEE. DOI: https://doi.org/10.1109/ICSC.2016.48 (2016)

Guma Lackshen is a PhD student at the School of Electrical Engineering. His interest includes NLP techniques, Linked Data, Big Data tools and technologies, and quality assurance in IT systems.

Valentina Janev is a Senior Researcher at the Mihajlo Pupin Institute, University of Belgrade, Serbia. Her interest includes business intelligence, decision support systems, Linked Data and Big Data tools, and applications of semantic technologies and W3C standards in different industrial domains. She has participated in many information systems projects for clients in Serbia and the region, as well as national and EU research projects. She serves as a Coordinator of the EU project LAMBDA, https://project-lambda.org/.

Sanja Vraneš, PhD is jointly appointed as the Director General of the Institute Mihajlo Pupin and as a Full Professor of Computer Science at the University of Belgrade. Her research interests include artificial intelligence, semantic web, linked data web, knowledge management, decision support systems, etc. From 1999 she has been engaged as a United Nations Expert for information technologies, and from 2005 as an expert evaluator and reviewer of EC Framework Programme Projects and H2020 projects.

Received: May 10, 2020; Accepted: November 01, 2020.