I.J. Information Technology and Computer Science, **2021, 6, 1-15**
Published Online December 2021 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2021.06.01

Linked Data: A Framework for Publishing Five-Star Open Government Data

Bassel Al-khatib
Syrian Virtual University, Damascus, Syria
Al-Sham Private University, Damascus, Syria
E-mail: t_balkhatib@svuonline.org, b.alkhatib.foit@aspu.edu.sy

Ali Ahmad Ali
Syrian Virtual University, Syria
E-mail: ali_85595@svuonline.org

Received: 03 July 2021; Accepted: 06 October 2021; Published: 08 December 2021

Abstract: With the increased adoption of open government initiatives around the world, a huge amount of governmental raw datasets was released. However, the data was published in heterogeneous formats and vocabularies and in many cases in bad quality due to inconsistency, messy, and maybe incorrectness as it has been collected by practicalities within the source organization, which makes it inefficient for reusing and integrating it for serving citizens and third-party apps.

This research introduces the LDOG (Linked Data for Open Government) experimental framework, which aims to provide a modular architecture that can be integrated into the open government hierarchy, allowing huge amounts of data to be gathered in a fine-grained manner from source and directly publishing them as linked data based on Tim Berners lee's five-star deployment scheme with a validation layer using SHACL, which results in high quality data.

The general idea is to model the hierarchy of government and classify government organizations into two types, the modeling organizations at higher levels and data source organizations at lower levels. Modeling organization's experts in linked data have the responsibility to design data templates, ontologies, SHACL shapes, and linkage specifications. whereas non-experts can be incorporated in data source organizations to utilize their knowledge in data to do mapping, reconciliation, and correcting data. This approach lowers the needed experts that represent a problem of linked data adoption.

To test the functionality of our framework in action, we developed the LDOG platform which utilizes the different modules of the framework to power a set of user interfaces that can be used to publish government datasets. we used this platform to convert some of UAE's government datasets into linked data. Finally, on top of the converted data, we built a proof-of-concept app to show the power of five-star linked data for integrating datasets from disparate organizations and to promote the governments' adoption. Our work has defined a clear path to integrate the linked data into open governments and solid steps to publishing and enhancing it in a fine-grained and practical manner with a lower number of experts in linked data, It extends SHACL to define data shapes and convert CSV to RDF.

Index Terms: Semantic web, linked data, open data, open government, open government data, linked open data, linked open government data.

1. Introduction

One vital step towards the realization of open government is the provision of open government data (OGD), such data has an important role in the data-driven society that aims to enhance government efficiency through data-driven participation, collaboration, and transparency. As well, this data has an economic impact stemmed from reusing it to support innovative services and applications provided for citizens and data-driven decision-making.

In fact, the public sector is one of the biggest sources of open data, e.g., toxic waste dumps, governmental spending, companies' registries, health facilities …etc. with the emergence of open government initiatives around the world, governmental agencies have released huge amounts of raw datasets on OGD portals, such as *data.gov* in the USA, *data.gov.uk* in the UK, *dados.gov.br* in Brazil [1, 2].

The published data were very useful for citizens to inspect and services to utilize the data that just exist in one structured dataset, however, the limitations begin when trying to reuse or integrate multiple structured datasets, that's due to heterogynous data formats and structures, inconsistent metadata and different names for the same entities [2, 3, 4]. Besides, the data was messy and in bad quality in many datasets, as they published as were produced without any

curation [5].

To rescue, the usage of semantic web and its guidelines that are known as linked data is hotly debated in the scholarly domain [6]. Until now, there is a lack of practical frameworks and platforms to publish linked open government data (LOGD), and existent approaches suffer from these Obstacles:

- Only the experts in semantic web technologies are involved in the process.
- They focus on converting data published on OGD portals rather than integrating such tools in the government hierarchy and publish data from source as linked data.
- They require many experts in the semantic web, which impedes convert many datasets.
- Data is converted as it, no validation is made, which leads to bad quality data that imped reuse and usefulness.

Our work aims to solve the aforementioned problems by providing the LDOG framework. The mental model behind it is making linked data the main format to publish government open data, in an effective and practical manner. The effectiveness may be achieved by hiding the complexity of the modelling and linkage processes from the lower levels of organizations(ex. branches of organizations, departments). Through providing a mechanism for higher levels organizations(ministries, organizations) to predefine data publishing templates, which may guide the publishing process and only requires knowledge in the nature of data to validate and correct it to increase quality. To test the framework we developed the LDOG platform that utilizes its modules to power graphical user interfaces that can be used by governmental bodies to publish their data. Then we employed our platform to publish multiple UAE datasets of different organizations. Then, on top of the published data. We developed a proof-of-concept app that shows the practicality of the framework and the power of linked data and encourages its adoption by governments.

The remainder of the paper is organized as follows. Related work is discussed in Section 0. Section 0 presents the background information related to open government and open government data, in addition to the basic principles of semantic web and linked data. In section 0, we introduce our framework LDOG. Then, we apply this framework to power the LDOG platform (Section 0). In Section VI we used the platform to convert multiple datasets and then build a proof-of-concept app. Finally, Section 0 concludes our work.

2. Related Work

In order to facilitate the publishing of high-quality linked open government data (LOGD) by government agencies; many researchers have investigated different approaches to automate many aspects of the process.

The goal of the UnBGOLD [1], which is a platform developed by the University of Brasilia in Brazil is to enhance the open datasets published by the university through an automated mechanism based on user interfaces to transform them into linked data. It is limited to inner datasets linkage through exact string matching and no way to link to outer datasets in linked open data cloud (LOD), hence the datasets are isolated and do not deserve the fifth star of Tim Berners lee's deployment scheme[1].

The work proposed in [7] presents the Datalift framework and platform, the framework introduces six steps required for transforming structured data into linked data. The platform is an implementation, which consists of independent modules, each one performing a task on the data from the previous module in sequential order. The most notable limitation is it requires solid knowledge in semantic web standards, so only experts can utilize it.

The approach in [3] takes a different perspective through depending on consumers to transform government data into linked data, this approach is known as "crowd-sourcing" which is a part of Government 2.0. In core, the process is achieved by extending OpenRefine[2], which is a powerful tool in cleaning and transforming data to support transforming structured data into RDF and providing provenance information represented in RDF according to Open Provenance Model Vocabulary (OPMV). The main problem with this method is it increases the entry costs for consuming LOGD.

The TWC LOGD [2] is an open-source infrastructure centered around the idea of converting structured data into raw RDF in an automated fashion without human intervention, after that experts may enhance it progressively by utilizing vocabularies and linkage to other datasets, this tool was used to transform a lot of *data.gov* datasets.

The European FP7 project Fusepool P3 [4] extends the linked data platform(LDP), which is a set of specifications recommended by W3C. Aims to provide full control over linked data repository such as creating new resources and get a representation of a resource …etc. By enabling a transformation of any source of data to linked data, through a chain of components that can be configured to address the need of each situation. Every component must provide an interface through a REST API which means it can be developed using any technology and hosted in a distributed manner.

The framework proposed in [8], draws general outlines and four steps to produce LOGD in Saudi. The first step is crawling governmental websites and collecting open data sets then cleaning them and preparing them to be modeled. In the second step, use vocabularies to generate RDF datasets as N-triple dumps. the third step focuses on linking resources to similar ones inside the dataset and in external LOD datasets and stores them in a triple store, the last step is providing public access to that data through a SPARQL endpoint.

[1] https://5stardata.info
[2] Previously known as GoogleRefine

Delivering Information of Government (DIGO) [9] is a general architecture that aims to solve the lack of semantic interoperability for the data published in web portals, through a mechanism to handle all sorts of data(structured, semi-structured, unstructured) and transform it to linked data to be reused and integrated by the public.

3. Background

In this section, we shortly describe the open government and open data. Then, we introduce the basic concepts behind the semantic web and linked data related to our work. These principles establish a framework to solve data heterogeneity and linkage issues and pave the way to open data on a global scale.

3.1. Open Government and Open Data

A. Open Government (OG)

In the mid-1990s and with the support of information and communication technology (ICT), governments around the world have started to deliver their services electronically via the internet, which is actually, what we called "electronic government" [10]. With the openness movement spread across the globe, such as open source, open data...etc. there were increased demands by citizens to open public sector information [11, 12]. which lead to issuing of many legislations target publishing government data: such as the European Union directive about Public Sector Information reuse in 2013 [5], Brazilian Law on Access of Information in 2011 [1].

Nevertheless, the term "open government" has been coined and has dramatically increased in popularity since Obama's memorandum 2009, known as Open Government Directive. It is based on three principles: **transparency** by utilizing ICT for providing information to citizens, enhance effectiveness and decisions making by engagements of citizens through **participation**, finally **collaboration** between governments bodies among them and with other outside parties, ranging from individuals to the private sector [11].

To fuel OG the open government data is a cornerstone for transparency and data-driven participation and collaboration.

B. Open Data (OD)

The Open Knowledge Foundation (OKF) defining OD as "Open data and content can be freely used, modified, and shared by anyone for any purpose", the only accepted additional requirements are those related to mention the provenance info [13].

C. Open Government Data (OGD)

Shortly, it is the governmental data published in accordance to open data requirements; however, the nature of this data requires additional restrictions, such as not open data that violates the privacy or specified by local laws as secrets. The golden rule here that all governmental data is open except some restricted datasets as defined by lawmakers and not vice versa [14].

D. Linked Open Government Data (LOGD)

LOGD is the open government data published in accordance with linked data principles as presented in Fig.1 [15]:

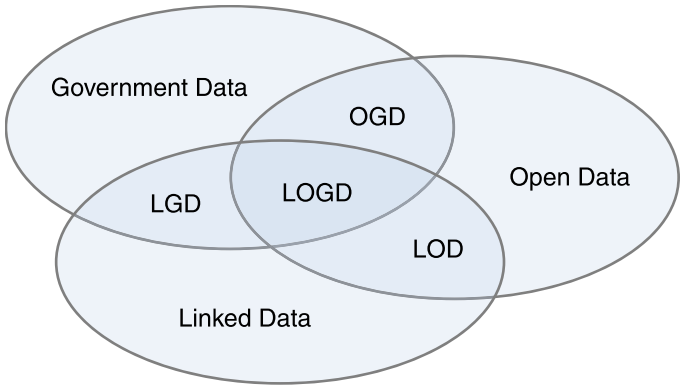

Fig.1. LOGD as the intersection of multiple domains [15]

3.2. Semantic Web and Linked Data

A. Semantic Web

The WEB (also known as the web of documents [6]) we use on our daily basis is a collection of billions of HTML pages interlinked using hyperlinks, people surf the web from page to page seeking for information. That is so good, but

what about machines? The answer is surely machines can go through web pages and retrieve their content. However, they are not able to understand the meaning of that content, due to the fact that HTML is designed towards representing information, not representing its semantics. Consequently, the software agents cannot use the web on behalf of us, maybe to find the nearest clinic and a reservation of an appointment.

To overcome this challenge, Tim Berners-Lee the inventor of the web decided to extends it to represent the data and its semantics that is what so-called "Semantic Web" [16] or Web of data [6].

The semantic web was built upon a set of standards recommended by w3c. The logic behind this is to identify each resource (ex. person, relation, place…etc.) using a global identifier namely the Uniform Resource Identifier (URI). Describe it using the Resource Description Framework (RDF), which is a graph data model consists of triples. the **subject** which is the resource we describe and the **predicate** which is a relation resource that links the subject to the third element named **object** which may be a literal or a subject for another triple and this where the actual linkage between triples in different sites around the world occur. The meaning of each element of the triple is provided using vocabularies and ontologies represented using the Resource Description Framework Schema (RDFS) and the Web Ontology Language (OWL). To query the web of data and manage datasets a set of standards known as SPARQL are introduced [17].

Lastly, in 2017 w3c added another standard to complete the semantic web stack, by providing the Shape Constraints Language (SHACL), to define the expected form of the RDF graph and a set of predefined validation rules with the ability to add new rules, such shapes may additionally be utilized to automatically generate user interfaces based on it [18].

The mature family of semantic web standards is seen by researchers as a great solution for data management difficulties such as interoperability between systems, integrating data from different sources, solve heterogeneity of data formats and metadata [9, 8, 4].

B. Linked Data Principles

Despite the fact that Semantic Web standards were mature and powerful, there was no clear path to realize the web of data, Trying to establish guidelines to bridge this gap, Tim Berners-Lee was introduced 4 steps to reach the linked data [19]:

1. Use URIs as names for things.
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL).
4. Include links to other URIs so that they can discover more things.

C. 5-star deployment scheme for Open Data

Follow up his previous efforts towards defining the principles of linked data, and to pave the way for different parties especially governments to upgrade their data to high quality linked data, Tim Berners-Lee placed a five-star deployment scheme, which open data can progressively track towards being a 5 star linked data [19]:

1. Available on the web (whatever format) but with an open licence, to be Open Data.
2. Available as machine-readable structured data (e.g. excel instead of image scan of a table).
3. As (2) plus non-proprietary format (e.g. CSV instead of excel).
4. All the above plus, Use open standards from W3C (RDF and SPARQL) to identify things, so that people can point at your stuff.
5. All the above, plus: Link your data to other people's data to provide context.

4. Framework

It is an experimental framework. Aims at providing an architecture for integrating linked data into open government hierarchy and defines clear modules for publishing data from source to public with a lower number of experts in linked data using a publishing pipeline based on predefined templates. we will first describe the architecture and then discuss the different modules.

4.1. Architecture and Schema

The ultimate goal of this framework is to be integrated into the government hierarchy, so publishing data from the source could be a routine task for open government. It harnesses the hierarchy to distinguish between two types of organizations: modeling organizations and data source organizations. The modeling organizations take a place at higher levels of the hierarchy (Cabinet, Ministries, Independent Agencies, Institutions) which are represented by departments in those organizations responsible about OGD, and the data source organizations take a place at the lowest levels (branches, departments) as illustrated in Fig.2.

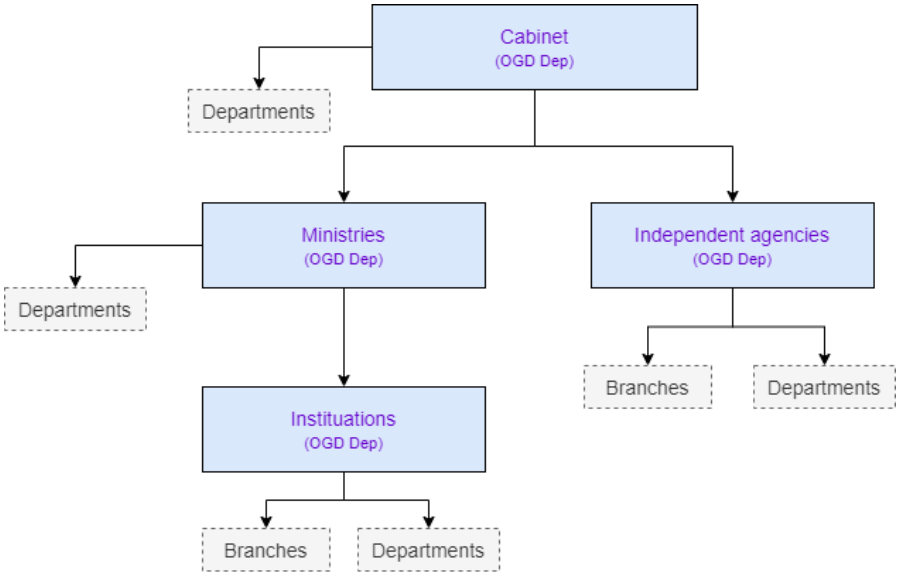

Fig.2. Lined-boxes represents modeling organizations, dashed-boxes represents data sources

The experts in the semantic web are only required at modeling organizations to develop ontologies, data templates, data shapes, and linkage specifications. In another hand, the data source organizations do not need any experts, normal employees can use data templates designed at higher organizations to publish their data and do the operations which need understanding the nature of data like mapping, reconciliation, and validation. The mental model behind this is to lower the number of experts and allow non-experts to contribute; furthermore, data could be accumulated in a fine-grained manner, which may increase the transparency of data.

To model this hierarchy and all the needed entities by the framework. we have developed two ontologies:

A. LDOG Ontology3

The objective of this ontology is to model the government hierarchy, organizations, user accounts, data shapes, data templates, batch imports…etc.

B. Conversion Ontology4

Government data is a trusted source and operations on it should be logged, so it can be inspected later. furthermore, it may be utilized to convert similar datasets and to ensure the transparency of the process. Thus this ontology aims at modeling the different operations that are implemented on data on the way to be five-star linked data.

The basic idea is to predesign everything needed to convert the dataset into LOGD by the modeling organization. in the form of what we call a "data template", which is ready to guide the process of publishing data, It consists of the following:

- data shape: which extends SHACL with more properties to automate the process as we will discuss in the next section.
- linkage specifications: define SILK framework specifications for linkage to other datasets.
- export target: which determines the organizations responsible for publishing their data based on this template.

In the data source organization. select the data template and upload the corresponding CSV to start the publishing pipeline that utilizes LDOG modules and employees' knowledge about the nature of data to produce LOGD.

Below is an example of a data shape for representing areas of the UAE, which we will explain its usage in the next section.

[3] https://github.com/ali-syria/ldog/blob/master/ontologies/ldog.ttl
[4] https://github.com/ali-syria/ldog/blob/master/ontologies/conversion.ttl

Data Shape 1. The data shape of representing a UAE area

```
@prefix sh: <http://www.w3.org/ns/shacl#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix ldog: <http://ldog.org/ontologies/2020/8/framework#> .
@prefix adm: <http://data.test/ontologies/administrative-division#> .
@prefix  : <http://data.test/shape/area-spape#> .

:AreaShape a sh:NodeShape ;
    a ldog:DataShape ;
    sh:closed true;
    sh:ignoredProperties (rdf:type rdfs:label) ;
    sh:targetClass adm:Area ;
    ldog:resourceIdentifierProperty adm:name ;
    ldog:resourceLabelExpression "{name}" ;
    sh:property [
        sh:path adm:name ;
        sh:name "name" ;
        sh:description "name of the area" ;
        sh:order 1 ;
        sh:datatype xsd:string;
        sh:minCount 1;
        sh:maxCount 1;
        sh:uniqueLang true ;
        sh:message "name field is required, string" ;
        ldog:normalizedBy ldog:Capitalize ;
    ] ;
   sh:property [
        sh:path adm:emirate ;
        sh:name "emirate" ;
        sh:description "emirate where area is located" ;
        sh:order 2 ;
        sh:class adm:Emirate ;
        sh:minCount 0;
        sh:maxCount 1;
        sh:message "invalid emirate" ;
   ] .
```

Across the process, we utilize a proper RDF serialization for each operation, for example, Turtle for data shapes, so experts can manually edit, JSON-LD for handling raw RDF data programmatically, N-triple for links to outer, which clarify the power of RDF.

Technically speaking, the framework has implemented as a package for the Laravel framework and tested using PHPUnit 8.5. It consists of the following modules:

4.2. Modules

A. URI Builder

The first and second principles of linked data focus on naming things or resources based on Unified Resource Locators (URLs). To establish a consistent and permanent URLs scheme, this module in the light of the recommendations proposed by W3C[5] and the UK Cabinet Office[6] , and the European Commission study on persistent URIs[7] defines a factory to produce the needed URLs to name the real resources, named graphs, data shapes, data templates ...etc.

The root URI is *http://{sector}.{domain}*, where the *sector* refers to the field of data (e.g. health, education, topography), it acts as a namespace and *domain* refers to an internet top-level domain name (e.g. data.sy, data.ae).

[5] https://www.w3.org/TR/cooluris/
[6] https://www.gov.uk/government/publications/designing-uri-sets-for-the-uk-public-sector
[7] http://philarcher.org/diary/2013/uripersistence/

The basic URI scheme for real resources is *http://{sector}.{domain}/{type}/{concept}/{reference}*, where *type* refers to the nature of the representation (e.g. resource, data, and page) and *concept* is the abstract class of the resource (e.g. city, river) where *reference* is the identifier of instance like (e.g. Damascus, Nile); based on it, three derived URI schemes are generated:

- Abstract resource URI

http://{sector}.{domain}/resource/{concept}/{reference}

- Resource HTML page

http://{sector}.{domain}/page/{concept}/{reference}

- Resource RDF representation

http://{sector}.{domain}/data/{concept}/{reference}

B. URI Dereferencer

As the third principle of linked data states that when someone dereference a resource URI you should provide useful information since linked data is for use by both humans and machines and taken into consideration that resource URI in linked data is actually HTTP URI. So based on HTTP content negotiation, we extend Laravel to check if the HTTP request for resource URI requires RDF or HTML based on the *Accept* header, so we direct the request through HTTP *303 Redirect* to the URI of the right representation. To retrieve the resource representation, this module sends a SPARQL *Describe* query to the triple store module.

As the cabinet OGD department is the highest modeling organization, it may design the most general shared ontologies related to administrative divisions and contact points and any others, making the possibility to utilize such vocabularies to improve the resource representation when viewed as HTML on browsers.

C. Triple Store

Concerning the persistence layer, there are many free and proprietary triple stores that implement SPARQL specifications, so we provide an abstraction layer based on the driver's concept, we provide a default driver for GraphDB[8] as a starting point.

The framework requires two separated triple stores, the first one is open, which citizens and software agents can access through a SPARQL endpoint and execute only read queries on government data, and the other is secured tailored to store passwords for user's accounts, with closed access by only authorized accounts to query.

D. Organizations Manager

In order to model the government hierarchy and distinguish between the data source organizations and modeling organizations and to associate data templates and batch imports to them. This module provides a factory to create such organizations.

E. Authentication

To provide secure authentication for users' accounts, we specify a secure triple store's repository connection to store the passwords hash and the hash algorithm based on the LDOG ontology. We extended Laravel with a Graph User Provider that supports authentication process using graph store instead of a traditional SQL database.

F. Ontologies Manager

Ontologies provide a shared formal vocabulary to talk about everything and inference new facts. This module provides a mechanism to import ontologies from any URI and provide metadata about the domain, namespace, preferred prefix, and description.

G. Shapes Manager

To enforce consistency at the data shapes level, we provide a mechanism to import data shapes written in Turtle syntax (the most human-friendly serialization of RDF) to the triple store. In addition to checking them against a basic shape using SHACL validator module to ensure that specific predicates exist at every *ldog: DataShape* and *sh: Property Shape* that may be used later to automate the batch import of data or designing input forms or to view this data.

[8] https://www.ontotext.com/products/graphdb

This includes that every shape is *sh: NodeShape* and *ldog: DataShape* at the same time, it should also have those predicates:

- *sh: target Class* to predetermine *rdfs:class* of each record of imported data.
- *ldog: resource IdentifierProperty* to predefine of which predicate of the resource is unique to generate the resource's URI reference part.
- *ldog: resource Label Expression* to define a pattern for generating *rdfs:label* of the resource, which may include multiple predicates.

At every *sh: PropertyShape* we need to force that each one has specific predicates:

- *sh: name* to specify a human-friendly name for each property, that can be used later by non-experts to map CSV column names to data shape properties. after that, the framework will use these mappings to map column names to the ontology's predicates.
- *sh: description* to describe the precise meaning and avoid ambiguity.
- *sh: min Count* to specify if the predicate is required or not.
- *sh: max Count* to specify the max cardinality of the predicate.
- *sh: order* to specify the order of the predicate, that may be useful to tidy viewing of dataset or sort the input forms.
- (*sh: datatype* to specify the data type of the value or *sh:class* to specify that value needs reconciliation with another dataset).
- *sh: message* to specify a human-friendly message in case of any violation.
- *ldog: normalized By* to specify normalization function for the cell value.

H. SHACL Validator

Providing valid data is an essential part of data usability and reusability for any purpose. To provide a layer to ensure consistency and correctness at the data level and shapes level. we used Apache Jena SHACL[9] command-line validator to validate imported data against data shapes and data shapes against the basic shape. In addition to data validation to ensure that every resource has a *rdfs:label* [10] which guarantees everything has a human-readable name that can be understandable when viewing resource as HTML and in reconciliation module for indexing purposes and other important goals as discussed further in [20].

I. Reconciliation

As the fourth principle of linked data suggests providing links to others' resources, the basic mechanism of doing that is to use the same resource's URI across datasets, to do that we need to replace the entity's name in the input dataset with the corresponding resource's URI in another dataset, we named this module "reconciliation" based on OpenRefine[11] terminology.

We believe that this operation requires human intervention, especially from data sources to reliably ensure accuracy and correctness of the replacement, because it will handle data coming from the data source and should be fully correct, and not as *owl:sameAs* that may automatically be calculated based on the measurements of similarity and added to the inputted data to provide context and "follow-your-nose" [17] links.

The basic idea is to index every resource using its *rdfs:label*, as the validation module ensure that every resource has it, then we could search for any term and get a list of the closest resources ordered by the score of similarity based on edit distance and syntactic similarities, like Levenshtein distance. Technically this module uses GraphDB Lucene for indexation and search. The reconciliation process requires *rdfs:class* of the target resource, in addition to the entity name. First, we check if there is a fully matched resource meaning that the edit distance between its *rdfs:label* and entity's name is zero(full-match), so we confidently replace it without any human intervention. Otherwise, this module aids the employee by providing an ordered list of candidate resources, that the employee go through each resource and dereference its URI to HTML for inspection, and select the proper one of them.

J. Normalization

In order to unify the data representation like date format (e.g. ISO 8601), capitalization, and any other special formats that ensure a consistent look and to avoid many pattern violation errors at the validation layer, we provide this module so it can automatically if possible transform value into a valid format or left as it. The normalization function is specified in the data shape on *sh:PropertyShape* shapes using *ldog:normalizedBy* predicate.

K. Outer Linkage

To gain the fifth star of Tim Berners Lee, we provide a mechanism to link to others datasets outside of government

[9] https://jena.apache.org/documentation/shacl/
[10] https://patterns.dataincubator.org/book/label-everything.html
[11] https://openrefine.org/

datasets through *owl: sameAs* and other special linkage links from other ontologies using the SILK framework. The modeling organization designs the SILK linkage specifications file, which may use different distance measures to state equality or other relations between resources in local datasets and any other resources in the LOD cloud that can be accessed through a SPARQL endpoint. Therefore, this module may provide a linkage to the outside and linkage to the global context.

L. Templates Builder

A template is a mechanism to accumulate the data shape, SILK framework linkage specs, data domain, export frequency, and export target, which are designed by experts in modeling organizations. We distinguish between two types: data collection template for non-temporal data and report template for temporal data like daily, monthly, yearly. Therefore, this module provides such a mechanism to construct a template resource according to LDOG ontology. The export target may be:

- All Branches, all Departments, all (branches and departments), and the modeling organization itself, which may be useful for importing taxonomies or categories (e.g. facility type, category of service).
- Specific branches or departments.

M. Batch Importer

The goal of our framework is to import huge amounts of data or batches according to a predefined data template. The contribution of this module in this process is to import the final converted JSON-LD file of the publishing pipeline and the conversion JSON-LD file (which log transactions performed on the dataset along the publishing pipeline) into named graphs in the triple store. in addition to metadata, related to the time range, publishing date, and the publishing organization.

We distinguish between two types: data collection batch Import for non-temporal data and report import for temporal data like daily, monthly, yearly.

N. Utilities

This tiny module includes retrieving basic LDOG enumerations or in OWL terms "Individuals" like data domains, data export targets of a data template, export frequency, term-resource match type, so it can be used to seed user interfaces or ease the retrieving them inside other modules.

O. Publishing Pipeline

This module aims to progressively enhance the inputted dataset by utilizing different modules to enhance data from one to five stars as illustrated in Fig.3.

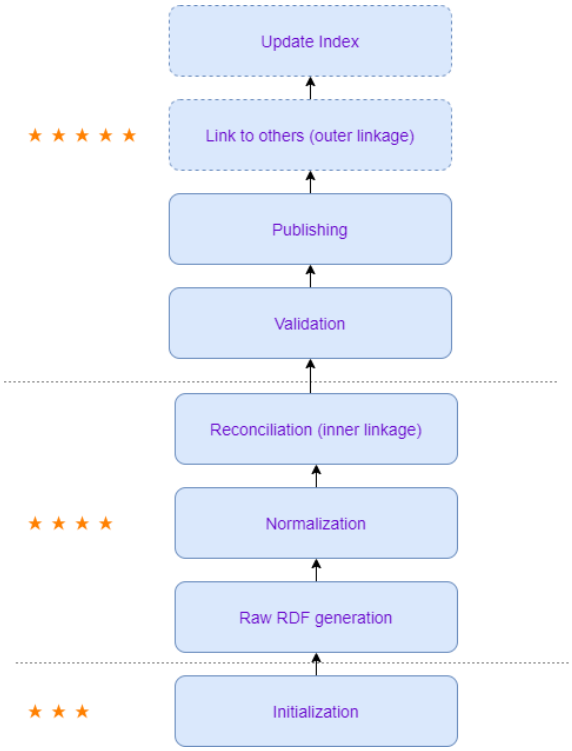

Fig.3. Lined-boxes represents foreground tasks, dashed-boxes represents background tasks

1. Initialization

At this step, the employee shall specify the data template and upload the corresponding CSV file to start the import process.

2. Raw RDF Generation

The first step is to parse the CSV file and construct a raw JSON-LD file. In fact, the tabular data as CSV has a triple architecture if we considered that each record represents a subject (resource), column names represent predicates, and cell values represent objects. To transform CSV to RDF based on the data shape. We only need human intervention to map between CSV column names and data shape properties that may be represented by their human-friendly *sh:name*, with extra explanation using *sh:description* to remove any ambiguity and aid the data source organization employee in this process.

The overall process is illustrated in the below pseudo-code and Data Shape 1

Algorithm 1 Algorithm for Raw RDF Generation

Input :
$C = \{R_1, R_2, \ldots R_n\}$ CSV file of records
$DS = \{NS, PS_1, PS_2, \ldots, PS_n\}$ Data shape which consists of Node Shape and Property Shapes
Output:
$J = \{T_1, T_2, \ldots, Tn\}$ JSON-LD file of RDF triples

EM = [sh:name => CSV column name] Employees in the data source organization map between CSV column names and the corresponding property shapes represented by its *sh:name* in the Data Shape.
M = [sh:path => CSV column name] mappings calculated based on *EM*, by replacing *sh:name* with the corresponding *sh:path*.
store mappings in the conversion file
NS = get Node Shape from *DS*
RC = get *sh:targetClass* from *NS*
RUP = get *ldog:resourceIdentifierProperty* from *NS*, its the unique predicate for the resource.
UC = extractUniqueCSVcolumnName(*M*,*RUP*)
foreach *R* **in** *C* **do**
  UniqueCellValue = *R*[*UC*]
  ResourceUri = build URI using *UniqueCellValue* as the reference part and *RC* as concept part and data domain part from the data template. By using the URI builder module.
  Resource = buildNode(*RC*, *ResourceUri*)
  foreach *m* **in** *M* **do**
    PropertyShape = extract the corresponding *sh:PropertyShape* from *DS*
    Predicate = get *sh:path* from *PropertyShape*
    CsvColumnName = get the corresponding column based on *m*
    CellValue = *R*[*CsvColumnName*]
    if *PropertyShape* **has** *sh:data_type* **then**
      DataType =*PropertyShape* [*sh:data_type*] the object of the corresponding triple is a typed literal
      Object = buildObjectLiteral(*CellValue*, *DataType*)
    else *PropertyShape* **has** *sh:class* **then**
      Object = *CellValue* the object of the corresponding triple is another resource. it will be replaced in the upcoming reconciliation step by the corresponding resource in another dataset.
    end
    addNewTriple(*Resource*, *Predicate*, *Object*)
  end
end

3. Normalization

Before the validation step and to avoid pattern violations errors like date format should be in ISO 8601, we automatically normalize the literal objects of the triples using the normalization module. The transformation functions for each object are predefined by data shapes using *ldog:normalizedBy* and no need for any human intervention. All normalization functions automatically transform value if possible or left as it, so it can be corrected manually later during validation.

4. Reconciliation

At this stage, the triples that their objects should be resources based on the corresponding data shape properties have *sh:class*, their raw object literals may be replaced using the reconciliation module, The goal of this step is the inner

linkage among government datasets, which we call the *inner linkage*.

5. Validation

The JSON-LD file coming from the previous step is validated using the SHACL validator module based on the data shape in the template. Taken into consideration that this file contains batch data, which means that the same errors may be repeated across the triples. For example, the use of "-" as a default value if a phone field is empty, which is an incorrect format for the phone number and will cause many validation errors.

To overcome this challenge, we will suspend the validation process on the first error that was detected, so the employee can correct the value or set it to empty if *sh:minCount:0* for that predicate in the data shape, and that can be interpreted by the framework by removing the entire triple. After that, the employee can apply this fix to the current triple or choose to bulk update all triples that have the same object for the same predicate as the current triple. Consequently, we can sharply decrease errors as validation progress.

6. Publishing

This step aims to publish the final converted JSON-LD file and conversion file into the open triple store using the batch importer module, so it will be available for reuse from the SPARQL endpoint. The reason for publishing the conversion file is for transparency and inspection purposes.

7. Linkage To Others Datasets

At this stage, the silk framework begins the linkage operation in the background to generate *owl:sameAs* links and other types of links according to SILK linkage specs defined in the data template. Then storing them into a dedicated named graph for outer links, so it can easily be inspected later and enhanced progressively and periodically.

8. Update Index

To ensure the proper work of reconciliation, we need in the background to update the index with the uploaded triples, so it can be discovered during the upcoming reconciliation processes by others.

5. LDOG Platform

As we previously illustrated, the framework provides a clearly defined set of modules each of one do a specific task in the process of publishing LOGD. To put our framework in action, we need to provide user interfaces for governmental organizations to access the power of the framework throughout government. Therefore, We developed the LDOG platform, which is a web application, built using the Laravel framework based on our LDOG Framework, and using Tailwindcss[12] and Alpine.js[13] for the frontend.

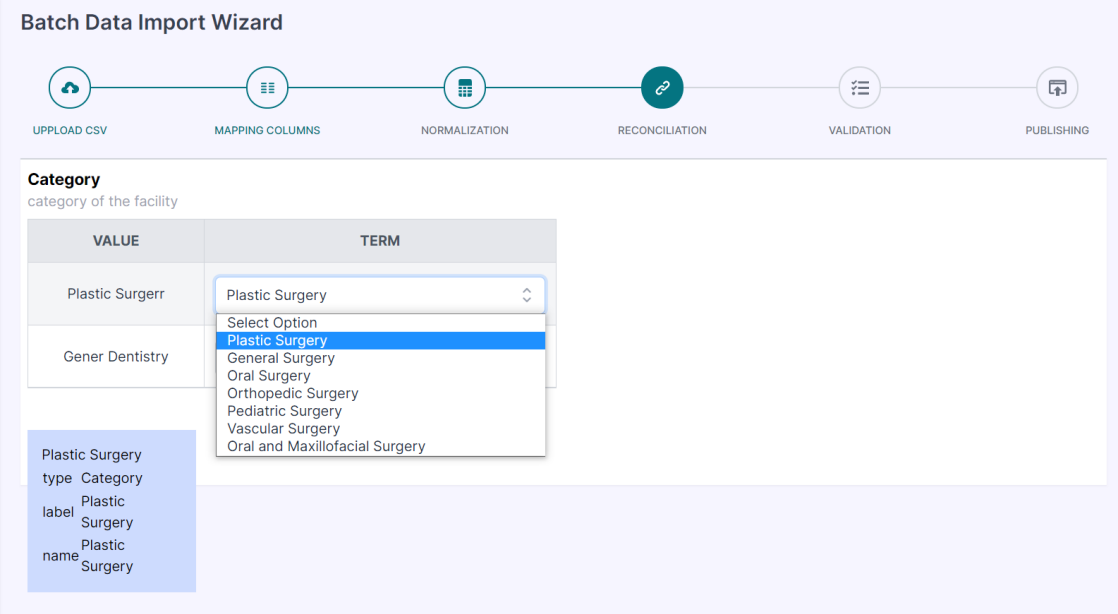

Fig.4. Reconciliation step between cell's value and the corresponding resource's URI, which represented by its rdfs:label.

[12] https://github.com/tailwindlabs/tailwindcss
[13] https://github.com/alpinejs/alpine

Its pages can be categorized into two groups. The first group is targeted at managing government organizations' accounts, import ontologies, and create data templates. While the second group targets publishing datasets using the publishing pipeline module of the framework. By providing human-friendly interfaces for mapping CSV columns to data shape properties, and doing reconciliation (Fig.4) and validation (Fig.5) steps as it requires human-intervention as we have seen.

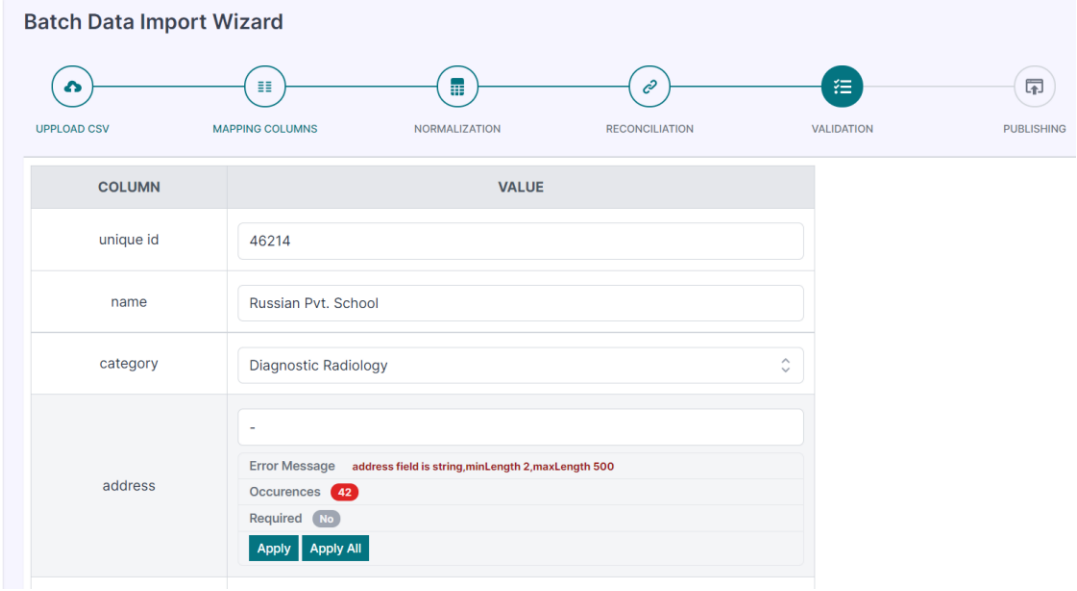

Fig.5. Validation step with the option to apply the fix to all records to speed the process

6. Results

In order to evaluate our framework in action, and show the benefits of five-star linked data to integrate data from multiple organizations, and solve heterogeneous formats and vocabularies issues. We have utilized the LDOG platform to convert multiple datasets from different UAE government agencies, published on *dubaipulse.gov.ae*, *bayanat.ae*, and *fcsa.gov.ae* to high-quality linked data. As shown in Table 1. We have linked the emirates of UAE to the corresponding resources in DBpedia using owl:sameAs links, in addition to Dubai's areas. To simulate the real-world usage, we used the following flow to publish the raw datasets:

1. First, we create an account for the Cabinet OGD department, using this account, we published general ontologies like Administrative Divisions & Location Description Vocabulary and Contact Point Vocabulary, which unify these vocabularies across the government, hence solve the heterogeneity of vocabularies issue. Then according to these vocabularies, we created data collection templates to publish the emirates and their areas, after that, we used them to publish the UAE's emirates & the areas of Dubai emirate, we also created accounts for the ministry of health and Prevention (MOHP), the ministry of interior (MOI) and The Federal Competitiveness and Statistics Centre (FCSC) (independent agency).
2. By using the health ministry account, we created the Statistics and Research Center (SRC) department, and then published the Covid-19 statistics vocabulary, after that, we used this vocabulary to make the Covid-19 daily statistics report template and assigned it to SRC. SRC used this template to publish its Covid-19 daily statistics per emirate immediately without any need for experts in linked data, emirate's name automatically reconciled to emirate's resource URI previously published by Cabinet in step 1. The health ministry also created another account for the institution of Health Authorities. Health Authorities published health facility vocabulary and based on it generated Health Facility data collection template, it has published Health Facility Categories, Health Facility Sub-Categories, and Health Facility Statuses, after that it created an account for Dubai Health Authority (DHA) branch and assign to it the Health Facility data collection template. The branch used this template to publish health facilities in Dubai without any experts and reconciling categories and sub-categories and statuses to resources published by the Health Authorities institution and correcting invalid data (ex. Facility's coordinates (90,90) which not in Dubai at all).
3. The ministry of Interior created an account for police forces institution; police forces published crime statistics vocabulary and crime types and created police crime statistics yearly report template and assigned it to the Dubai police branch. Dubai's branch used this template to directly publish their data as linked data without the need for any experts, but only their knowledge in data nature.
4. The Federal Competitiveness and Statistics Centre (FCSC) (Independent Agency) published price indices

vocabulary, created “monthly percentage change in consumer price index by emirate” report template, and published their data according to it.

This mechanism unifies vocabularies by centralizing them, solves the heterogeneity issue of published data by depending only on RDF as the only format to publish data. Increase transparency through publishing data from source and in a fine-grained manner not as just accumulated statistical reports, ease linkage by depending on higher organizations like ministries to register categories and taxonomies that provides a central registry of trusted data and reconcile to them across government. Lowering the linked data experts by only requesting them in higher levels. Whereas non-expert employees in data source organizations use predefined data publishing templates and their knowledge in the data nature to publish data routinely. The framework also assists in leveraging data quality by validating datasets in the source.

Table 1. Organizations datasets

Agency	*Datasets*
Dubai Health Authority (DHA)	Active health facilities[14]
Federal Competitiveness Statistics Authority (FCSC)	Consumer Price Index[15]
Ministry of Interior (MOI)	Crimes By Type Of Crime And Emirate[16]
Ministry of Health and Prevention (MOHP)	COVID-19 Daily Updates[17]

After that, we built a proof of concept app based on SPARQL queries as shown in Fig.6., The idea is a mash-up app that starting from an area of Dubai, the linkage to DBpedia and integrating datasets from disparate organizations to provide data about the description and population of the area and Dubai. Locations of health facilities in that area from DHA, The statistics of the latest 10 days of COVID-19 confirmed cases in the UAE. In addition to the statistics of the latest 6 months percentage change in consumer price in Dubai. Lastly, old statistics from 2007 about crime in Dubai. This may be very useful to anyone who needs to rent a house or book a hotel room in that area.

Fig.6. POC App

7. Conclusions

Open government data plays a significant role in the realization of open government aspirations and a valuable resource of economic growth. However, publishing raw government data can impede this bright ambition due to bad quality and linkage issues. The current approaches are utilizing linked data techs to solve those issues. Nevertheless,

[14] https://www.dubaipulse.gov.ae/data/dha-location/dha_sheryan_facility_detail-open
[15] https://fcsa.gov.ae/en-us/Pages/Statistics/Statistics-by-Subject.aspx#/%3Ffolder=Economy/Prices/Consumer%20Price%20Index
[16] http://data.bayanat.ae/en_GB/dataset/crimes-by-type-of-crime-and-emirate
[17] https://apps.fcsc.gov.ae/Covid19Updates/api/api/Covid19UAE/Download/En

The existent frameworks of publishing are limited to experts usage only, with no validation tier and it is based on converting already published data rather than integrating them into the government hierarchy and publishing data immediately as linked data in a fine-grained approach, which lead to more transparency and providing a huge amount of high quality trusted data.

Our LDOG framework solves the aforementioned problems and making the publication of high-quality LOGD a responsibility and the routine task of governmental agencies. With the lowest number of experts and engagement of non-experts in the process to utilize their knowledge in the nature of data to do non-technical tasks, such as mapping, reconciliation, and validation based on predefined data templates, which are designed by experts in higher levels of the government hierarchy.

We put our LDOG Framework under investigation by developing the LDOG platform, which utilizes LDOG framework modules to convert multiple datasets that were published by the UAE government and building a proof of concept app on top of it, to demonstrate the benefits for governments and citizens.

Our framework is experimental and needs more work to be production-ready. Despite that, it draws solid modules and steps to make publishing linked open government data more practical and efficient. The most notable lack of LDOG is a clear and transparent mechanism to update and delete resources.

References

[1] L. C. B. Martins, M. C. Victorino, M. Holanda, G. Ghinea, and T.-M. Grønli, "UnBGOLD: UnB Government Open Linked Data: Semantic Enrichment of Open Data Tool," in *Proceedings of the 10th International Conference on Management of Digital EcoSystems - MEDES '18*, 2018, pp. 1–6, doi: 10.1145/3281375.3281394.

[2] L. Ding *et al.*, "TWC LOGD: A portal for linked open government data ecosystems," *J. Web Semant.*, vol. 9, no. 3, pp. 325–333, 2011, doi: 10.1016/j.websem.2011.06.002.

[3] F. Maali, R. Cyganiak, and V. Peristeras, "A Publishing Pipeline for Linked Government Data," in *The Semantic Web: Research and Applications. ESWC 2012. Lecture Notes in Computer Science*, vol. 7295 LNCS, Springer, Berlin, Heidelberg, 2012, pp. 778–792.

[4] L. Selmi and N. Alessia, "Fusepool P3: A Linked Data Platform for Open Government Data," *Electron. Gov. Electron. Particip. Jt. Proc. Ongoing Res. Proj. IFIP WG 8.5 EGOV ePart 2015*, pp. 101–108, 2015, doi: 10.3233/978-1-61499-570-8-101.

[5] A. Vetrò, L. Canova, M. Torchiano, C. O. Minotas, R. Iemma, and F. Morando, "Open data quality measurement framework: Definition and application to Open Government Data," *Gov. Inf. Q.*, vol. 33, no. 2, pp. 325–337, 2016, doi: 10.1016/j.giq.2016.02.001.

[6] B. Todesco, B. Blume, A. Zancanaro, J. L. Todesco, and F. Gauthier, "Linked open government data research panorama," *IC3K 2013; KEOD 2013 - 5th Int. Conf. Knowl. Eng. Ontol. Dev. Proc.*, pp. 278–285, 2013, doi: 10.5220/0004548402780285.

[7] F. Scharffe *et al.*, "Enabling Linked Data Publication with the Datalift Platform," *Twenty-Sixth AAAI Conf. Artif. Intell. - Work. Semant. Cities*, vol. WS-12-13, 2012.

[8] Afnan M. AlSukhayri, Muhammad Ahtisham Aslam, Sachi Arafat, Naif Radi Aljohani, "Leveraging the Saudi Linked Open Government Data: A Framework and Potential Benefits", International Journal of Modern Education and Computer Science(IJMECS), Vol.11, No.7, pp. 14-22, 2019.DOI: 10.5815/ijmecs.2019.07.02

[9] A. L. Machado and J. M. Parente de Oliveira, "DIGO: An Open Data Architecture for e-Government," in *2011 IEEE 15th International Enterprise Distributed Object Computing Conference Workshops*, 2011, pp. 448–456, doi: 10.1109/EDOCW.2011.34.

[10] P. Milić, N. Veljković, and L. Stoimenov, *Smart Technologies for Smart Governments*, vol. 24, no. July. Cham: Springer International Publishing, 2018.

[11] B. W. Wirtz and S. Birkmeyer, "Open Government: Origin, Development, and Conceptual Perspectives," *Int. J. Public Adm.*, vol. 38, no. 5, pp. 381–396, 2015, doi: 10.1080/01900692.2014.942735.

[12] A. Corradi, L. Foschini, and R. Ianniello, "Linked data for Open Government: The case of Bologna," in *2014 IEEE Symposium on Computers and Communications (ISCC)*, 2014, pp. 1–7, doi: 10.1109/ISCC.2014.6912473.

[13] OKF, "The Open Definition - Open Definition - Defining Open in Open Data, Open Content and Open Knowledge," *Opendefinition.org*. [Online]. Available: http://opendefinition.org/. [Accessed: 13-Feb-2021].

[14] M. Kaschesky and L. Selmi, "Fusepool R5 linked data framework," in *Proceedings of the 14th Annual International Conference on Digital Government Research*, 2013, pp. 156–165, doi: 10.1145/2479724.2479748.

[15] B. S. Hitz-Gamper, O. Neumann, and M. Stürmer, "Balancing control, usability and visibility of linked open government data to create public value," *Int. J. Public Sect. Manag.*, vol. 32, no. 5, pp. 457–472, 2019, doi: 10.1108/IJPSM-02-2018-0062.

[16] B. Tim Berners-lee, J. Hendler, and O. Lassila, "The Semantic Web A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities," *Sci. Am.*, vol. 284, pp. 34–43, 2001.

[17] T. Heath and C. Bizer, "Linked Data: Evolving the Web into a Global Data Space," *Synth. Lect. Semant. Web Theory Technol.*, vol. 1, no. 1, pp. 1–136, Feb. 2011, doi: 10.2200/S00334ED1V01Y201102WBE001.

[18] "Shapes Constraint Language (SHACL)." [Online]. Available: https://www.w3.org/TR/shacl/. [Accessed: 28-Feb-2021].

[19] "Linked Data - Design Issues." [Online]. Available: https://www.w3.org/DesignIssues/LinkedData.html. [Accessed: 03-Jul-2020].

[20] B. Ell, D. Vrandečić, and E. Simperl, "Labels in the web of data," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2011, vol. 7031 LNCS, no. PART 1, pp. 162–176, doi: 10.1007/978-3-642-25073-6_11.

Authors' Profiles

Bassel Al-khatib is the web sciences master director at the Syrian Virtual University and the head of Artificial Intelligence department at Information Technology Faculty at Damascus University. He holds PhD degree in computer science from the University of Bordeaux-France, 1993. Dr. Alkhatib supervises many PhD students in web mining, and knowledge management. He also leads and teaches modules at both BSc and MSc levels in computer science and web engineering in Syrian Virtual University, Damascus University, and Al-Shem Private University.

Ali Ahmad Ali is a web sciences master student at Syrian Virtual University. He has a Bachelor degree in information technology engineering, Tishreen University, Lattakia, Syria, 2016. He has been working as a Web Developer since 2015.

How to cite this paper: Bassel Al-khatib, Ali Ahmad Ali, "Linked Data: A Framework for Publishing Five-Star Open Government Data", International Journal of Information Technology and Computer Science(IJITCS), Vol.13, No.6, pp.1-15, 2021. DOI: 10.5815/ijitcs.2021.06.01