



UNIVERSITY OF BELGRADE
FACULTY OF ELECTRICAL ENGINEERING

Guma Abdulkhader LAKSHEN

A framework for analysis and quality assessment
of big and linked data

Doctoral Dissertation

Belgrade, 2021



УНИВЕРЗИТЕТ У БЕОГРАДУ
ЕЛЕКТРОТЕХНИЧКИ ФАКУЛТЕТ

Гума Абдулкахдер ЛАКШЕН

Окружење за анализу и оцену квалитета
великих и повезаних података

докторска дисертација

Београд, 2021

Предлог састава Комисије за преглед и оцену докторске дисертације (предлаже Ментор):

1. _ Професор Др Сања Вранеш _____
(ментор)
2. _ Професор Др Бошко Николић _____
(члан комисије са изабраног модула кандидата - препорука, није обавезно)
3. _ Др Валентина Јанев, виши научни сарадник _____
(члан комисије који није у радном односу на ЕТФ-у)

Defense/датум одбране: _____

Abstract

Linking and publishing data in the Linked Open Data format increases the interoperability and discoverability of resources over the Web. To accomplish this, the process comprises several design decisions, based on the Linked Data principles that, on one hand, recommend to use standards for the representation and the access to data on the Web, and on the other hand to set hyperlinks between data from different sources.

Despite the efforts of the World Wide Web Consortium (W3C), being the main international standards organization for the World Wide Web, there is no one tailored formula for publishing data as Linked Data. In addition, the quality of the published Linked Open Data (LOD) is a fundamental issue, and it is yet to be thoroughly managed and considered.

In this doctoral thesis, the main objective is to design and implement a novel framework for selecting, analyzing, converting, interlinking, and publishing data from diverse sources, simultaneously paying great attention to quality assessment throughout all steps and modules of the framework. The goal is to examine whether and to what extent are the Semantic Web technologies applicable for merging data from different sources and enabling end-users to obtain additional information that was not available in individual datasets, in addition to the integration into the Semantic Web community space. Additionally, the Ph.D. thesis intends to validate the applicability of the process in the specific and demanding use case, i.e. for creating and publishing an Arabic Linked Drug Dataset, based on open drug datasets from selected Arabic countries and to discuss the quality issues observed in the linked data life-cycle. To that end, in this doctoral thesis, a Semantic Data Lake was established in the pharmaceutical domain that allows further integration and developing different business services on top of the integrated data sources. Through data representation in an open machine-readable format, the approach offers an optimum solution for information and data dissemination for building domain-specific applications, and to enrich and gain value from the original dataset. This thesis showcases how the pharmaceutical domain benefits from the evolving research trends for building competitive advantages. However, as it is elaborated in this thesis, a better understanding of the specifics of the Arabic language is required to extend linked data technologies utilization in targeted Arabic organizations.

Keywords: Linked Data, Open data ecosystems, Drug management applications, methodology, Quality assessment, Quality dimensions, Tools, Drugs Application, Application: Arabic Datasets

Scientific area: Electrical engineering and computer science

Narrow scientific area: Software engineering

Апстракт

Повезивање и објављивање података у формату "Повезани отворени подаци" (енг. *Linked Open Data*) повећава интероперабилност и могућности за претраживање ресурса преко *Web*-а. Процес је заснован на *Linked Data* принципима (*W3C*, 2006) који са једне стране елаборира стандарде за представљање и приступ подацима на *Вебу* (*RDF*, *OWL*, *SPARQL*), а са друге стране, принципи сугеришу коришћење хипервеза између података из различитих извора.

Упркос напорима *W3C* конзорцијума (*W3C* је главна међународна организација за стандарде за *Web*-у), не постоји јединствена формула за имплементацију процеса објављивање података у *Linked Data* формату. Узимајући у обзир да је квалитет објављених повезаних отворених података одлучујући за будући развој *Web*-а, у овој докторској дисертацији, главни циљ је (1) дизајн и имплементација иновативног оквира за избор, анализу, конверзију, међусобно повезивање и објављивање података из различитих извора и (2) анализа примена овог приступа у фармацеутском домену.

Предложена докторска дисертација детаљно истражује питање квалитета великих и повезаних екосистема података (енг. *Linked Data Ecosystems*), узимајући у обзир могућност поновног коришћења отворених података. Рад је мотивисан потребом да се омогући истраживачима из арапских земаља да употребом семантичких веб технологија повежу своје податке са отвореним подацима, као нпр. *DBpedia*-јом. Циљ је да се испита да ли отворени подаци из Арапских земаља омогућавају крајњим корисницима да добију додатне информације које нису доступне у појединачним скуповима података, поред интеграције у семантички *Веб* простор.

Докторска дисертација предлаже методологију за развој апликације за рад са повезаним (*Linked*) подацима и имплементира софтверско решење које омогућује претраживање консолидованог скупа података о лековима из изабраних арапских земаља. Консолидовани скуп података је имплементиран у облику Семантичког језера података (енг. *Semantic Data Lake*).

Ова теза показује како фармацеутска индустрија има користи од примене иновативних технологија и истраживачких трендова из области семантичких технологија. Међутим, како је елаборирано у овој тези, потребно је боље разумевање специфичности арапског језика за имплементацију *Linked Data* алата и њихову примену са подацима из Арапских земаља.

Кључне речи : Повезани подаци, Отворени екосистеми, апликације за управљање медикаментима, методологија, процена квалитета, димензије квалитета, Софтвер, Апликација за медикаменте, Апликација: Арапски скуп података

Научно област: Електротехника и рачунарство

Уже научно област: Софтверско инжењерство

Acknowledgements

*Firstly, I would like to thank my supervisor, Prof. **Sanja Vraneš**, and my great assistant supervisor Dr. **Valentina Janev**, for their continuous and endless support, their continuous invaluable suggestions, prolonged guidance, and patience throughout the past six years. Without their guidance, I would have never completed this Ph.D.*

Many thanks to everybody at the School of Electrical Engineering and Mihajlo Pupin Institute for their help and collaboration.

My gratitude and appreciation are also to the Libyan ministry of higher education for their trust by presenting the financial support of my Ph.D. scholarship

*Last but not least, I would like to thank the soles of my parents and my beloved family; my wife **Feaza**; my sons **Montaser**, **Zakaria**, and **Mohammad**; my daughters **Huda**, **Saida**, **Aya**, and **Nausayba** for their tremendous understanding, encouragement, and patience throughout my study.*

Guma Lakshen

Contents

Chapter One - Introduction	2
1.1 Background and Motivation.....	3
1.2.1 Challenges with Big and Open Data	3
1.2.2 Reuse of Drug information from the Web for building innovative applications4	
1.2.3 Arabic Language Contents on the Web – Facts and Challenges.....	5
1.3 Research Goals and Challenges	6
1.3.1 Research Challenges Related to development Linked Data Applications.....	6
1.3.2 Research Goal	7
1.3 Research Questions	8
1.4 Contributions.....	8
1.4.1 Systematic Literature Review	8
1.4.2 Conceptual Methodology for Linked Data Quality Assessment	9
1.4.3 Consolidating the Arabic Open Drug Data	9
1.4.4 ALDDA-QA Framework.....	9
1.5 Thesis Outline	9
Chapter Two – The Semantic Web Space and Big Data	12
2.1 Introduction.....	13
2.1.1 The Semantic Web Stack	14
2.1.2 From Web 1.0 to The Internet of Things (IoT).....	16
2.3 Linked Data.....	19
2.3.1 Linked Data Principles and the 5-star Open Data Model	19
2.3.2 Linked Open Data Best Practices.....	21
2.4 Semantic Web Languages	22
2.4.1 RDF/XML.....	23
2.4.2 Resource Description Framework Schema (RDFs).....	24
2.4.3 Taxonomies and Thesauri	26
2.4.4 Web Ontology Language (OWL)	27
2.4.5 The SPARQL Query Language	28
2.5 Big Data and the Web – a state of the art.....	29
2.5.1 Big Data Definitions and Characteristics	29
2.5.2 Importance and Benefits of Big Data.....	33
2.5.3 Big Value Created from Big Data.....	34
2.5.4 Challenges of Big Data	36
2.5.5 Tools and Technologies of Big Data.....	36
2.6 Characteristics of a Modern Data Ecosystem	38
2.7 Summary	39
Chapter Three – Quality Issues of Linked data Ecosystems	42
3.1 Introduction.....	42
3.2 Generic use-case in a Linked Data Ecosystem	44
3.2.1 Challenges	45
3.2.2 Components of an Arabic Linked Open Drug Data Ecosystem	46
3.3 Quality of Linked Open Data – State of the Art	49
3.3.1 Data Quality life-cycles	49
3.3.2 Data Quality Issues	51
3.3.3 Data Quality Problems Classification	53
3.3.4 Quality Dimensions of Linked Open Data	54
3.3.5 Challenges Facing Linked Data Quality Dimensions	58

3.3.6 Data Quality Dimensions Classifications Schemes	59
3.4 Linked Data Quality Assessment Methodologies - Comparison	60
3.4.1 Definitions.....	60
3.4.2 Strategies and Techniques for Linked Data Quality Assessment	61
3.4.3 Comparison of Previous and Related Works	62
3.4.4 Comparison of Linked Data Quality Assessment Frameworks	64
Open-source.....	65
3.5 A Conceptual Methodology for Linked Data Ecosystems Quality Assessment	68
3.5.1 Proposed Methodology	68
3.5.2 Selection of Data Quality Dimensions for quality evaluation rules.....	72
3.6 Summary	75
Chapter Four – “TowardS Solution Development” for CONSOLIDATION of Arabic Linked Drug Datasets	77
4.1 Introduction.....	77
4.2 Towards the development of a Knowledge Graph.....	77
4.2.1 Using the RDF data model.....	77
4.2.2 Ontologies and Knowledge Graphs	79
4.3 Selection of LOD and Arabic Linked Drug Datasets	81
4.3.1 Existing Arabic Drugs-related Datasets on the Web	81
4.3.2 Linked Open Drug Data LODD.....	82
4.4 Interlinking and enhancing Arabic Drugs Datasets	84
4.4.1 Selection of Linked Open Data Tools.....	86
4.5 ALDDA Piloting methodology and QA framework development ...	88
4.6 Selection and Implementation of Data Quality Assessment Measures	90
4.7 Validation of the ALDDA Approach.....	93
4.7.1 Implementation of the Arabic Linked Drug Data Application (ALDDA)...	94
4.8 Summary	103
Chapter Five – Results and Analysis	105
5.1 The Arabic Linked Drug Data Application Quality Assessment Architecture	105
5.2 Consolidating Arabic Open Drug Data	107
5.3 Quality Assessment of Arabic DBpedia	113
5.4 Summary	115
Chapter Six - ConclusionS and Future Directions	117
6.1 Analysis of the Linked Data Lifecycle	117
6.2 Quality Analysis of Integrated Open Data	118
6.3 Proposal for further development of quality assessment tools	119
Bibliography	122

Figures

Figure 1: Integrating public and private datasets.....	4
Figure 2: The World's 10 most Influential languages.....	5
Figure 3: The four layers of the Semantic Web pyramid.....	14
Figure 4: The Semantic Web Stack Standard.....	15
Figure 5: Web 1.0 illustration.....	16
Figure 6: Web 2.0 illustration.....	16
Figure 7: The transition from Web of documents to the Web of data.....	17
Figure 8: Web 3.0 illustration.....	17
Figure 9: Linked Open Data cloud (2020) and LODD [337].....	21
Figure 10: W3C best practices to publish a dataset as LOD.....	22
Figure 11: A simple taxonomy relationship tree.....	26
Figure 12: A drugs thesaurus relationship tree.....	26
Figure 13: The original 3Vs of big data.....	31
Figure 14: Big Data Knowledge Discovery.....	34
Figure 15: Big Data Value chain.....	35
Figure 16: Five-ways supporting value creation from big data.....	35
Figure 17: Generic Linked data ecosystem use-case.....	45
Figure 18: Modern data ecosystem for Arabic Linked Open Drug Data.....	46
Figure 19: Generic data quality life-cycle.....	50
Figure 20: Data Quality Dimensions [Source DAMA, 2013].....	58
Figure 21: Methodology for Assessing Linked Open Data Quality.....	69
Figure 22: A Flow Chart of the Methodology for Assessing Linked Open Data Quality.....	70
Figure 23: RDF graph example.....	78
Figure 24: RDF extended example.....	78
Figure 25: Example of connections between data using nodes and edges. Source [Exploring Knowledge Graphs for COVID-19 Drug Discovery CAS].....	80
Figure 26: A diagram of the LODD datasets [337].....	83
Figure 27: Piloting methodology phases.....	88
Figure 28: A novel linked data methodology with a focus on quality assessment.....	93
Figure 29: Data mapping from Arabic datasets.....	96
Figure 30: The ALDDA: Drug class.....	97
Figure 31: The reconciliation process based on actCode, genericName, ChemicalSubstances, and Drug synonyms.....	99
Figure 32: Knowledge graph visualization and querying.....	102
Figure 33: Quality Assessment Framework - Simplified illustration.....	106
Figure 34: Two SPARQL queries indicating the interlinkable drug's data.....	107

Tables

Table 1: Web development stages since 1995	18
Table 2: Initial Datasets in the LOD cloud as in 2007	20
Table 3: Big data characteristics and their relation to quality	32
Table 4: Research issues and application domains in data quality literature and EU research projects.....	43
Table 5: Data Lake vs Data Warehouse	48
Table 6: Data quality problems classification in data sources.....	54
Table 7: List of research articles discussed data quality dimensions	55
Table 8: Most frequently used data quality dimensions along with their criterion.....	56
Table 9: Comparison of linked data methodologies.....	62
Table 10: The proposed Linked Open Drug Data methodology	64
Table 11: Comparison of existing data quality assessment frameworks and tools.....	65
Table 12: Selected Arabic open drug datasets	85
Table 13: Selected LODD Datasets to be interlinked with Arabic Drug datasets.....	85
Table 14: RDF Transformation tools.....	86
Table 15: Linked data application phases and detailed steps	88
Table 16: Data quality assessments functional forms.....	91
Table 17: The ALDDA merged property file after mapping	96
Table 18: Data Quality dimensions relevant for quality assessment of Arabic DBpedia (*Specific to DBpedia, **Specific to Arabic DBpedia)	101
Table 19: Big Data challenges of and implemented functionalities related to quality.....	119
Table 20: Comparison of open-source quality assessment tools according to several attributes	120

Acronyms

3V's	Volume, Velocity, and Variety
ACID	Atomicity, Consistency, Isolation, Durability
AI	Artificial Intelligence
DAMA	International Data Management Association
DARPA	Defense Advanced Research Projects Agency
DAWG	Data Access Working Group
DBMS	database management systems
DQV	Data Quality Vocabulary
HCLS IG	HealthCare and Life Sciences Interest Group
HTTP	Hyper Text Transfer Protocol
HIQA	Health Information and Quality Authority
ICT	Information and Communications Technology
IoT	Internet of Things
IRI	Internationalized Resource Identifier
LOD	Linked Open Data
LODD	Linked Open Drug Data
MENA	Middle East and North Africa
MIT	Massachusetts Institute of Technology
ML	Machine Learning
ODM	Open Data Movement
OIL	Ontology Interchange Language
OWL	Web Ontology Language
RDF	Resource Description Framework
RDFs	Resource Description Framework schema
RIF	Rule Interchange Format
ROI	Return of Investment
SWLs	Semantic Web Languages
TDQM	Total Data Quality Management
URI	Uniform Resource Identifier
W3C	World Wide Web Consortium
WWW	World Wide Web
XML	Extensible Markup Language

Chapter one

Introduction

CHAPTER ONE - INTRODUCTION

The World Wide Web (WWW) impacted our world significantly, as it changed our view of how we share information by permitting its users to publish documents in a generally open global information space. “*Web documents*” can contain hypertext links that allow users to navigate additional documents to discover additional related information, enhancing the value of the original datasets. The WWW leads innovation domains by diverting the information community from the concept in which data-owners had a dominant data repository such as a database, to a Web-dominated community in which various data sources need to interrelate and interoperate, in a way that gives a fully integrated view of distributed information [1]. Furthermore, data-driven institutions are starting to aggregate data from multiple data sources, rather than just relying on their own (proprietary) data silos, this aggregation could then be fuelled back to enterprise “data lakes”, in an attempt to develop a big data ecosystem.

The quality of the information provided could differ as information providers have *different knowledge levels, diverse views of the world, different intentions and objectives, and diverse anticipated outputs*. The significance of achieving and preserving data with a high-quality standard is widely recognized by practitioners and researchers. Based on its influence on businesses, data quality is commonly regarded as a valuable asset. Data with low-quality levels almost certainly can have catastrophic and far-reaching costs for a business, such as *poor decision-making and missed business prospects*, since the provided data might not reflect the clear picture of the circumstances [2][3][4]. Enabling good quality information that can precisely answer complicated queries and lead to effective decision-making remains one of the most significant challenges facing the data life cycle. To achieve it, efficiently extracting and integrating information from diverse, distributed, and heterogeneous data sources are required to generate such a good quality knowledge. Therefore, dissimilar to traditional desktop applications, Web applications require the ability to handle the distributed features of the Web, the issues that arise from mutual information sharing, and especially to deal with heterogeneity and uncertainty flexibly and efficiently [1].

Zaveri et. al (2016) published a survey that observed an extensively varying data quality ranging from comprehensively curated datasets to crowdsourced and data extracted of relatively low quality [5]. Gathering and publishing big volumes of structured data is viewed as an optimistic phase in the right direction. However, the quality of the gathered and published data still raises a serious obstacle towards the complete utilization of big data applications at a large scale. A critical challenge to data quality is the dynamic nature of linked data where data can witness a rapid change and flop to imitate changes in the real world, thus information becoming obsolete. *Zaveri et al.* (2014) identified the challenges to the Linked data as openness, information diversity, unbounded dynamic set of autonomous data sources; and publishers. Also, providing semantic links, detecting datasets quality, and making the information explicit pose new challenges [6]. Hence, before the information is utilized to perform a specific task, information quality should be regularly and carefully measured against a task-specific criterion.

A survey conducted by Experian Information Solutions¹ (2016) showed that 83% of the participants state that poor data quality affected their business objectives and 66% report that poor data quality has had a negative influence on their organization in the last 12 months. Also, KPMG² 2016; Forbes³ Insights 2017, reported that 84% of CEOs are worried about the quality of the data they use for decision making [7][8]. IBM Research (IBM Big Data and Analytics Hub⁴) in 2016 estimates that in the U.S., the total annual costs resulting from poor data quality is estimated at \$3.1 trillion [9].

The Semantic Web strength is its ability to interlink datasets using the available structured meta-data. If users can create links between different datasets, they can boost the value of datasets with relevant information from the interlinked dataset (e.g., if a webpage provides the user with drugs prices, the user would not mind having relevant data from the DrugBank webpage for further knowledge and information from DBpedia for enhancing lingual information). To achieve this interlinking, datasets are required to be published to present their meta-data, whereas, most datasets are published without their meta-data. A transformation process is required to structure and submit this relevant meta-data. Presently, a single way to perform this transformation does not exist as well, and the tools to perform automatic interlinking between datasets to publish and enrich them are lacking.

1.1 Background and Motivation

1.2.1 Challenges with Big and Open Data

The big growth of the Internet and the process of data generation cannot be handled by existing technologies and poses challenges for technology providers.

The majority of data generated online is mostly in text format; that can be easily understood and processed automatically. Additionally, data duplication on the Web is an extra major challenge (about 30% of the total volume of the data on websites is redundant⁵) and requires quality assessment mechanisms. As a result, the information retrieved with Web queries is not accurate and does not refer to the data required in the query, but rather to documents that contain the data. Security and privacy issues pose a threat since sharing information on social media networks could result in the misuse of individual information.

In the last twenty years, the emerging technology trends (Big Data, Linked Data, semantic technologies) foster new approaches to the design and implementation of enterprise knowledge management systems that are based also on the reuse of open datasets from the Web. Hence, this thesis proposes a method to use Semantic Web techniques, explains the way they can operate together, to transform an isolated dataset into a Linked Data dataset bearing in mind data quality issues.

¹ [Business Solutions | Experian](#)

² [KPMG US LLP - KPMG United States \(home.kpmg\)](#)

³ <https://www.forbes.com>

⁴ [The IBM Big Data and Analytics Hub – Government Aggregator](#)

⁵ [How Common is Duplicate Content? - Raven \(raventools.com\)](#)

1.2.2 Reuse of Drug information from the Web for building innovative applications

Drug information is scattered widely on the Web in a distributed manner, due to the regulations adopted by governments, and organizations independently curate the drug data available on the Web by local institutions in each country, allowing the data to exist in disparate languages, with uneven structure and format located in diverse places on the Web of data. This attitude hinders and limits the harvesting of value and insights from the scattered drug information. To overcome this shortcoming, the interlinking and integration of big and/or open heterogeneous data into a comprehensive dataspace for developing data-driven applications are becoming more demanding. There is substantial information about drugs and pharmaceuticals obtainable on the Web. Data sources cover a wide range of research zones as medicinal chemistry results, drugs impact on gene expression, drugs result in clinical trials. Linked Data best practices adoption paved the way for the Web into the global data space, interlinking data from various domains, such as scientific publication, proteins, genes, clinical trials, and drugs, etc. Linked Data principles adoption permitted data publishers to provide structured interlinked data with additional open datasets on the Web in an efficient manner to terminate their isolation. The development of a global dataset of pharmaceutical information requires defining methodological guidelines and developing specialized tools for creating Linked Data in the drug domain usable on a universal scale.

In Arab-speaking countries, the existence of large drug datasets is limited and mostly prepared in the English language as it is the academic language for doctors and pharmacists. The available datasets are mostly not open and not updated regularly. The diversity of data structure is well observed as there are no unification and coordination between related organizations in those countries, leaving their data isolated and inefficiently valued. The general end-users in the Arab countries are mostly unacquainted with foreign languages and only speak their mother tongue, i.e. Arabic. Creating, Interlinking, and consolidating diverse Arabic drug data in one data lake is aiming at unifying dataset structure, fixing quality issues, providing useful additional information missing (e.g., DrugBank) in the original datasets, and giving abstracts and additional available information in another knowledge graphs (e.g., DBpedia).

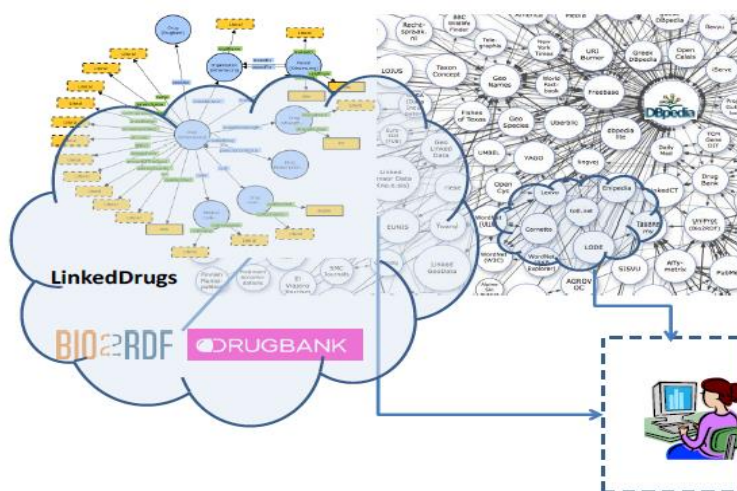


Figure 1: Integrating public and private datasets

1.2.3 Arabic Language Contents on the Web – Facts and Challenges

The Arabic language is the official language of the twenty-two Arab countries in the Middle East and North Africa (MENA region) spoken by more than 422 million, according to World Population Review 2020⁶, and the most spoken language in the Semitic language group⁷. It is the liturgical language of 1.8 billion Muslims around the world and is one of the six official languages of the United Nations.

Arabic is one of the world's ten most influential languages according to the World's 10 most influential Languages [332]. The Arabic language is the 5th most influential language in the world⁸ (see Figure 2). According to Wikipedia⁹, Arabic is the 4th language used on the Web with 237.4 million users representing 5.2% of total users worldwide. Despite the widespread of the Arabic language, the situation is dimmer regarding the Arabic language content in WWW, it is < 3%¹⁰; the situation is even worse concerning open data, linked data, and open drug linked data.

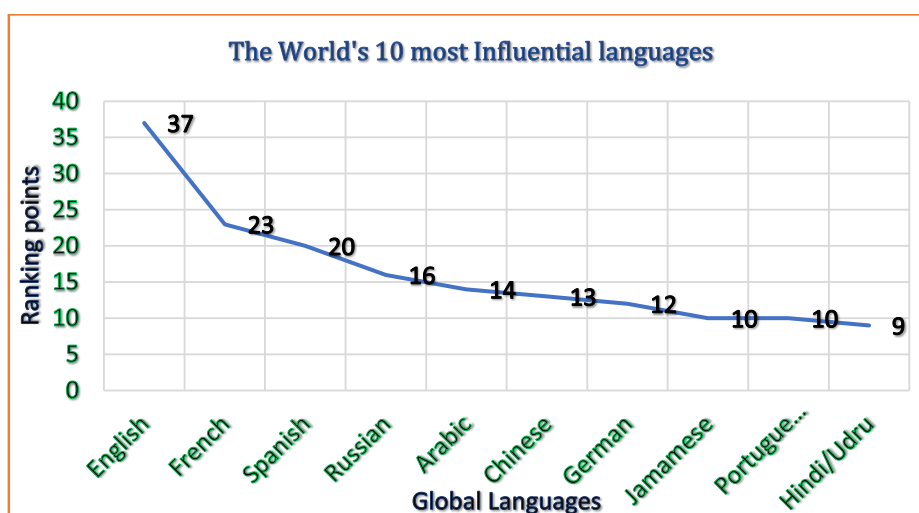


Figure 2: The World's 10 most Influential languages

This is due to the Arabic language having a set of specialties that made it a tough language and may hinder the development of Semantic Web tools for it. Among these specialties, its complex morphological, grammatical and Semantic features, since it is an extremely inflectional and derivational language [333]. Furthermore, the Arabic Language has no capitalization property, which directly affects and complicates the identification of the Arabic Named Entities, i.e., harder to identify proper names, acronyms, and abbreviations. Moreover, the Arabic Language is tremendously ambiguous, due to several reasons such as the vowelization feature of the Arabic Language, which causes ambiguity when it is not included, and this is a usual case, Polysemous (multiple words meaning), which are words that share the same spelling and pronunciation but have different meanings [64][334]. An additional issue that affects Semantic

⁶ <https://worldpopulationreview.com/country-rankings/arab-countries> [Accessed 15-10-2020]

⁷ [Semitic languages | Definition, Map, Tree, Distribution, & Facts | Britannica](#)

⁸ <https://thecareercafe.co.uk/blog/10-most-influential-languages-in-the-world-which-languages-will-make-you-most-employable/> [Accessed 15-10-2020]

⁹ [Languages used on the Internet - Wikipedia](#)

¹⁰ <https://www.khaleejtimes.com/naton/dubai/arabic-content-is-less-than-3-on-world-wide-Web>

Web tools processing for Arabic script is the problem of encoding since different encodings for Arabic script exist on the Web [335]. Thus, it is necessary to develop tools to help users to exploit the content of the Web in this language. This limitation of Arabic content encourages us to enrich the Arabic language user to gain value and insight by utilizing Semantic Web technologies by interlinking Arabic data with other datasets such as in English languages.

1.3 Research Goals and Challenges

Over the last few years, increasing deployment of Linked Data applications has been initiated as a standard framework to publish interlinked structured data on the Web, which enables public and private organizations users to fully employ a big volume of data from manifold domains that did not exist previously. The advances in the Web technologies encouraged the Web content producers to plunge it with huge amounts of information, but with less quality as *fake news, unreliable statistics, inconsistent data, irrelevant information, and misleading information*.

Hence, the principal research question of this doctoral thesis concerns how to design and implement the Linked Data lifecycle to fully leverage the potentials of semantic technologies for building Linked Data applications on top of open datasets from Arabic countries (used as an illustrative use case), while mitigating the risks of integrating poor quality datasets. This thesis proposes a method of creating a Linked Data dataset out of existing Excel datasets, created by four different organizations in four different Arabic countries. Within this converting process, the interlinking phase is not straightforward, since the datasets are currently not published as Linked Data. This means that the relevant meta-data is not available and computers are unable to understand its structure.

There is numerous existing dataset that can be transformed into Linked Data datasets to enrich and increase the datasets usage opportunities. Once the transformation process is completed, computers will be able to access the relevant meta-data and capable of understanding the underlying data structure afterward, the data needs to be published. This whole process generates difficulties and numerous decisions need to be made.

1.3.1 Research Challenges Related to development Linked Data Applications

Many causes prevent the Semantic Web from achieving its full potential yet, such as the unavailability of sufficient linked data to work with, and the unwillingness of converting existing datasets to linked data format before the development of applications that will use it. The Semantic Web is confronted with many challenges and concerns, the major ones are: 1) *availability of content*; 2) *ontology availability and evolution*; 3) *scalability*; 4) *visualization to reduce information overload*; 5) *multilingualism*; and 6) *stability of Semantic Web languages* [14]. Other challenges include: i) *identifying Semantic annotations by data providers without expectations of an instant bonus*; ii) *resolving Semantic heterogeneity issues that emerge when diverse data owners create semantically optimized representations of data, and iii) quality issues decline the level of data users' trust* [1]. Also, data quality assurance is still faced with many

challenges, such as defining the measurement methods to identify the level of data quality, especially because it is highly dimension dependent [15]. These challenges, among others, cause poor data quality, including imperfect measurement and assessment methodologies. Therefore, the efficacy of the Semantic Web is highly reliant on the existence of machine-readable data, i.e., Linked Data and its related concepts and services [12]. Though, the data quality assurance process is still not advanced enough to have standard management methods to deal with poor data.

1.3.2 Research Goal

The following specific research goal was defined:

“Develop an approach to create Semantic Web applications by transforming Arabic Drug datasets into RDF, enriching this RDF by interlinking entity URIs of some diverse datasets from the Linked Data cloud, and publishing the resulting RDF as Linked Data taking into consideration data quality issues.”

Despite the existing data quality challenges and the importance of solving data quality issues, fairly little research considers the correlation between data quality dimensions and data quality frameworks. This thesis aims to tackle the aforementioned Semantic Web utilization problems by proposing a Linked Data generation methodology (Conceptual Methodology for Linked Data Ecosystems Quality Assessment) for increasing the data quality of consolidated datasets and improving their interoperability.

The thesis objective is to propose a set of complementary techniques and corresponding implementations that enable the Semantic Web’s adoption. Each one addresses a part of the envisaged high-quality Semantic enhancement and integration in the form of Linked Data from semi-structured heterogeneous data. The thesis also aims to investigate if data quality can be assessed within data quality dimensions to see how it can be measured, what corrective steps can be taken, and how data quality can be observed for continuous improvement. To attain this aim, understanding the various data quality dimensions is crucial, and it should be stated how data quality dimensions are affected, what interrelationships contribute to the data quality, and what interrelationships lead to decreasing quality of data.

Thus, this research develops a comprehensive assessment framework that identifies the quality issues and integrates the available datasets, and interrelates with acquired information in a form of a semantic data lake. The uppermost goal is to facilitate high-quality Linked Data generation independently of the available original data. How each part contributes is attested by evaluating the execution’s performance and validation’s results, and applying it in different use cases, among others open government, scientific research, industry, and generic domain knowledge.

1.3 Research Questions

This thesis originates its essence from the disciplines of, *Semantic Web, information technology, information management, business management, quality management, and performance management.*

Taking the drug industry and drug management as an example, this thesis was motivated by the following research questions:

- What is the Semantic Web and What is Linked Data?
- Which languages and techniques are used in developing Linked Data applications?
- What are Linked Open Data Ecosystems?
- What is the quality of data in the Open Data Ecosystems, e.g., the Arabic drug datasets? How can data quality dimensions be used to assess the quality of the dataset?
- What phases are required in transforming a dataset into Linked Data? How can the automatic interlinking process between entity URIs in Linked Data be validated? What are the benefits of integrating openly available data sources (e.g., DBpedia and DrugBank) into the existing business value chain, and what are the flaws of this approach? How can business intelligence services (e.g., a search operation) be applied on top of a Semantic drug data lake?

1.4 Contributions

In this thesis, our main contributions to the Linked open data ecosystem is (1) the Systematic Literature Review; (2) the proposed methodology for Linked Data Ecosystems Quality Assessment; (3) the consolidated Arabic Linked open drug dataset; and (4) development of a framework that includes data quality measurement methods tailored for Arabic datasets.

1.4.1 Systematic Literature Review

- Conduct a Systematic Literature Review highlighting notable articles related to Linked data frameworks. Most of the literature review conducted throughout the thesis was in the form of comparisons. We reviewed the development stages of the Web (Web1-Web3), big data characteristics from the original 3Vs to the 10Vs, data quality life-cycles, Linked, data quality dimensions, Open Data methodologies, best practices, and data quality assessment frameworks. In particular, we conduct a Systematic Literature Review to highlight the Linked Data quality dimensions needed for processing datasets from Arabic countries.

1.4.2 Conceptual Methodology for Linked Data Quality Assessment

- Provide a categorized summary of linked open drug data methodologies based on literature reviews and previous practical use cases. Based on the review of previous methodologies, we proposed a data quality assessment methodology, embedded within the linked drug data process, that allows different quality aspects of the integrated resource to be assessed and improved iteratively.

1.4.3 Consolidating the Arabic Open Drug Data

- Evaluate existing software tools and systems for tabular data cleaning and transformation frameworks that can solve most of the common data quality issues. Based on this, we proposed and implemented a framework for data cleaning and transformation operations that includes data quality measurement methods having in mind the needs of organizations from the pharmaceutical industry that does business with Arabic countries. Introduce the ALDDA piloting methodology used for the transformation process and its validation.

1.4.4 ALDDA-QA Framework

- From the review, and based on the requirement of the Arabic datasets, we selected the accuracy, consistency, and relevancy dimensions as these three dimensions represent the major problems that require validating. The three defined quality dimensions are studied intensively together with several related quality factors that are specific to the data integration context. The definitions of these quality criteria and factors are capable of forming the quality requirements from different categories of users.

1.5 Thesis Outline

This thesis is structured as follows:

Chapter 2 introduces the concepts of big and open data ecosystems which constitute the Semantic Web and its languages, big data ecosystems its components and characteristics, benefits and importance, knowledge discovery from big data and value chain, big data issues from the Web perspective. The chapter familiarizes the reader with the essentials of the Semantic Web.

In **Chapter 3** the relation between linked and big data is discussed from the point of view of quality assessment, especially linked open data quality dimensions and data quality life-cycle. Data quality issues and challenges are reviewed based on existing literature. In this chapter, based on previous works and reviews, we propose a generic data quality life-cycle and a methodology for assessing linked open data quality and the processes required for assessment. Also, we selected

the dimensions accuracy, consistency, and relevancy and their calculation methods that going to be focused on our Arabic datasets.

Chapter 4 introduces the proposed framework for analysis and quality assessment and gives a detailed discussion of its components which embeds quality assessment within the data integration process. We studied the previous related works and compared them to our proposed methodology. We introduce the ALDDA piloting methodology used for the transformation process and its validation. In this chapter, we study the quality dimensions relevant to Arabic DBpedia.

Chapter 5 presents the results and findings of the research. We introduce our data integration methodology which embeds quality assessment within the DI process.

Chapter 6 discusses thesis contributions and identifies some areas of future work.

Chapter Two

The Semantic Web Space and Big Data

CHAPTER TWO – THE SEMANTIC WEB SPACE AND BIG DATA

Humanity is witnessing the information age, where data is generated in huge volumes at a rapid rate; as a consequence, a large number of diverse processes and devices that produce data such as log files, sensors, transaction records, mobile devices, etc., and the high velocity with which data are created. These huge data volumes “*Big Data*” inherits diverse characteristics (such as variety, velocity, complexity, multi-format, multi-channel, and so on) that cannot be managed properly by traditional computer systems [17][18]. At the same time, computational works and storage costs dropped sharply, which, along with the increase in data sizes, motivated researchers to lay the foundations of big data technologies [19].

Big data technology can analyze and cross-reference large-sized data, and extract useful knowledge, insights, and value [20]. These technologies paved the way for private and public sector stakeholders to gain value and insight from outside data as well as their own [21]. *Tim Davies (2011)*, introduced the idea of fostering an *Open Data Ecosystem* to help identify and evaluate possible strategies that government and non-government Open Data Initiatives ODI¹¹ can adopt in seeking the realization of the promised benefits of open data [22].

Open data ecosystems are expected to bring many advantages, such as stimulating citizen participation and innovation. Big Data ecosystems development and implementation in organizations is a complicated process that comprises many technological aspects as well as management of policies and people [19]. Also, implementing Big Data and Semantic Web systems in organizations involves the collaboration and coordination of different stakeholders, as well as the synchronization and execution of many tasks and activities. Open data ecosystems are expected to bring many advantages, such as stimulating citizen participation and innovation.

In this chapter, we will discuss the Semantic Web space through the following sections:

Section 2.2 discusses the concept of the Semantic Web starting from defining the Semantic Web stack and its layers, via the development phase of the Web from web1.0 to Web 3.0 over the past decades, to the current Semantic Web challenges.

Section 2.3 Discusses the Linked Data Principles and the 5-star Open Data Model and how the linked drug's data evolved within the LOD cloud and illustrating the LOD best practices and the technologies which support Linked Open Data

Section 2.4 Illustrates the Semantic Web languages including RDF, RDF/XML, RDFs, OWL, ontologies and knowledge graphs, taxonomies and thesauri, RIF, and SPARQL.

Section 2.5 Discusses the interconnection between the Semantic Web and the Big Data via analyzing definitions and characteristics and discussing the varieties of V's (3Vs, 5Vs, 10Vs) and how big data influenced the attitude of knowledge acquiring and value gaining via Semantic

¹¹ [The Open Data Initiative](#)

Web. This section ends up by discussing the big data challenges and listing the tools and technologies utilized by big data.

Section 2.6 points to the characteristics of a modern data ecosystem.

2.1 Introduction

The term “*Semantics*” is a terminology utilized by linguists and logicians to describe the study of meaning. Semantics explores how perceptions or scenarios are codified into a language through a particular system of symbols to simplify communication between entities. In the case of information systems, entities are computers. *Semantic Web*, as a term was first coined by Foucault (1966) in *The Order of Things* book [23]. The World Wide Web Consortium¹² (W3C) Defined Semantic Web as “*The Semantic Web offers a common framework that permits data to be reused and shared through applications, organizations, and community boundaries. The Semantic Web is a cooperative endeavour led by W3C with collaboration from a large number of researchers, scholars, and industrial partners*”¹³.

Tim Berners-Lee the inventor of the Web, stated at the first international WWW conference at CERN¹⁴, Geneva, in September 1994, “*to a computer, then, the Web is a flat, boring world devoid of meaning . . . this is a pity, as documents on the Web describe real objects and imaginary concepts, and give particular relationships between them*” [24]. Tim Berners-Lee et al. (1998) defined the Semantic Web as “*it isn’t a discrete Web but an extension of the existing one, in which information is given well-defined sense, better-enabling computers, and users to work cooperatively....*” [25]. Berners-Lee was aiming at expanding the Web beyond hypertext into something more *Semantic*, as follows:

- I. *Sharing data and facts* rather than displaying the text content of a Web page.
- II. *Developing a technology stack to support “Web of data”* instead of the “*Web of documents*”.
- III. *Providing services that enable computers to process meaningful tasks and to upgrade systems that can support reliable interactions over the network.*

The Semantic Web adds value to the current Web by bringing structure to the content of Web pages, making it understandable to software tools, so enabling computers to understand Web pages similar to humans [12]. To achieve this, information is fed with well-defined meaning, enabling software tools to comprehend and process the information and carry out sophisticated tasks for humans, rather than just displaying data [12]. The exponential and uncontrolled growth of the Web makes the process harder and more complex. The increasing complexity is caused by the acceleration and virtually uncontrolled growth of the Web, leading to the ultimate need for more intelligent software agents that are becoming more and more crucial [26]. Another complexity is created by the emergence of new Web technologies, allowing further integration of more complex data sources. The Semantic Web development

¹²The World Wide Web Consortium (W3C) is an international community that develops open standards to ensure the long-term growth of the Web. Available at <https://www.w3.org/>

¹³ [Semantic Web \(umbc.edu\)](http://www.umbc.edu)

¹⁴ [Home | CERN](http://www.cern.ch)

stages, started ever since the creation of the internet, can be viewed in terms of a pyramid of layers (see Figure 3).

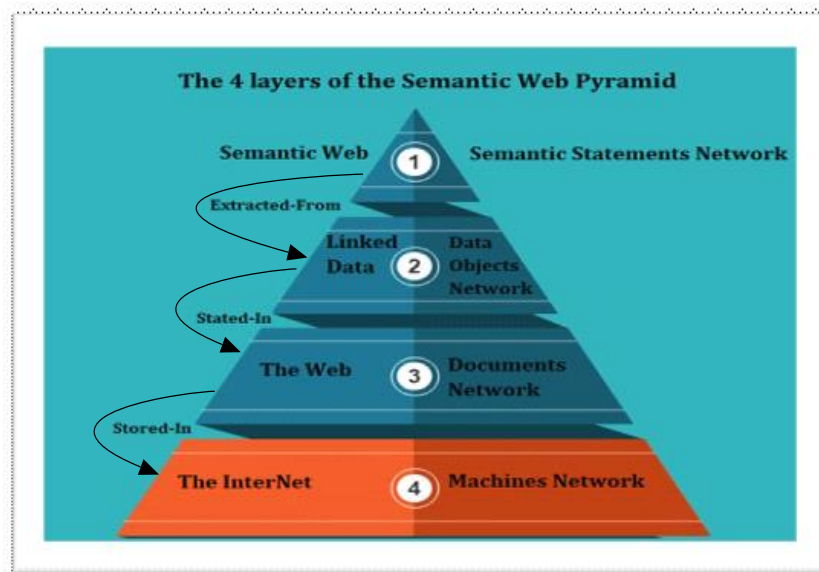


Figure 3: The four layers of the Semantic Web pyramid

To confront Web expansion, a transition from the existing Web which is also known as the *Web of Documents*, currently in use, to the Semantic Web, also known as the *Web of Data* becomes more demanding. The basic distinction between the two Web's is that the existing Web treats the entire document as their initial object, meaning that they are not capable to provide context to data [27]. Software tools cannot understand data meaning; therefore, it is unable to discriminate between the relevant and irrelevant portions of the document, similar to human anticipation and intuition. On the other hand, the primary objective of the Semantic Web is to concentrate on the resources (or their description). This resource is identified by a unique identifier called the *Uniform Resource Identifier/Internationalized Resource Identifier (URI/IRI)* which identifies the resource and a Web document describing the resource [28].

2.1.1 The Semantic Web Stack

Semantic Web stack is observed, in Scientific literature, as *Semantic Web cake* which describes the architecture of the Semantic Web, and demonstrates that the Semantic Web isn't a new technology rather than it's the extraction of traditional hypertext Web [29]. Most of these technologies such as RDF, SPARQL, OWL are represented in the Semantic Web Stack, which illustrates the architecture of the Semantic Web, as shown in Figure 4. W3C, the architectural designer of the Semantic Web since 1999, proposed a set of standards to technically back up this movement. Practically, the standards are built following a "*layer cake*" structure where standards are constructed hierarchically on top of lower ones.

Many layers exist and each layer benefits from the technologies of the lower layer and has a well-defined function in the architecture. Almost all of these layers are already implemented.

Undeniably, the languages and protocols that achieve their functions already existed or were designed and created to meet the specifications of each layer.

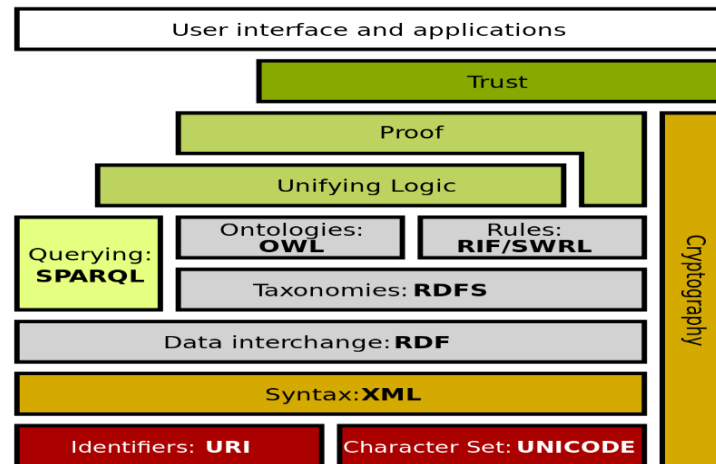


Figure 4: The Semantic Web Stack Standard

[Source: https://commons.wikimedia.org/wiki/File:Semantic_Web_stack.svg]

The top layer, i.e., the user interface layer, permits humans to utilize Semantic Web applications. The bottom layers of the stack embody the elementary hypertext Web technologies (URI, Unicode, XML). Layers at the top of the stack consist of technologies that are not yet standardized by the W3C or are still in the recommendation stage. The unifying Logic, Proof and Trust layers have not been implemented yet. They will build on top of each other to enable the identification and validation of information collected through RDF data. The cryptography layer ensures and verifies that the statements from the Semantic Web originated from a trusted source.

The middle layers contain the implemented and standardized Semantic Web technologies. The Data interchange layer represents the Resource Description Framework (RDF). RDF Schema (RDFS) is a model for RDF data, providing a data-modeling vocabulary. The OWL layer represents the Web Ontology Language which is an RDF-based language. The SPARQL (Protocol and RDF Query Language) is a protocol and a language that allows users to query the published RDF data on the Web [30]. A more detailed description of the Semantic Web Stack can be found in the book “*Handbook of Semantic technologies*” (2011) by *Domingue et al.* [31].

Two landmarks of the Semantic Web were the first W3C recommendation of the primary RDF standard in 1999 determining the fundamental data model [32], and the seminal paper published by *Berners-Lee et al.* (2001), where the authors outlined their vision for the Semantic Web [12]. The Semantic Web initiative’s (a W3C initiative) main idea was to enable linkage between remote data entities so that several aspects of information become available at once. The Semantic Web mainly depends on the dereferencing concept, where identifiers are used to represent entities and are therefore to scroll from one entity of information to another [33].

2.1.2 From Web 1.0 to The Internet of Things (IoT)

Semantic Web is the latest version in a series of Web versions Web1.0, Web2.0, and Web3.0 that helps stakeholders extract the precisely needed information, using machines instead of humans. Nowadays, the Web is used to achieve three major important tasks, i.e., *data searching*, *data combining*, and *data mining*. Semantic Web uses techniques where search tasks are performed based on the word meaning and what the user is thinking about [34].

In 1989, *Tim Berners-Lee* had a new vision for the Internet when he started developing the WWW. He envisioned a read/write Web-enabling the Web to be more interactive. In the 1990's *Lee* invented **Web 1.0** (also known as the *search Web* or the *static Web*), which was a basic and abstract read-only Web (see Figure 5). Web 1.0 enabled users to visualize the information without posting anything [35][36].

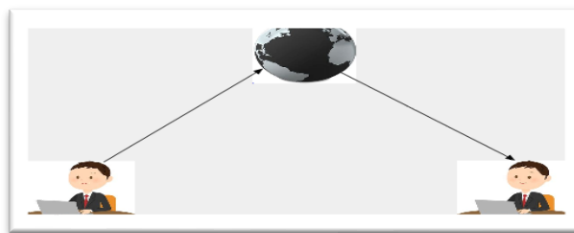


Figure 5: Web 1.0 illustration

Web 1.0 uses HTML, HTTP, URI technologies, and other protocols like XHTML, XML, and CSS. Web 1.0 also, combined technologies between server and client such as PHP, ASP, CGI, JSP, and PERL. The server uses JavaScript, VBscript, and flash on the client [35]. The main drawbacks are low speed and frequent site refreshing is needed whenever Web pages are modified, besides, it is a one-way direction platform, which means that a user cannot modify or post a Web page [19][20].

In 1999, **Web 2.0** was invented by *Darcy Di Nucci*, later in 2004, Web 2.0 (also known as the *sharable Web* or the *dynamic Web*) was popularized by *Dale Dougherty* and *Tim O'Reilly* at the media Web 2.0 conference (see Figure 6) [39]. Web 2.0 employed internet technologies enabling them to become bi-directional (more interaction with less control). Web 2.0, is a read-write network application that allows users to share and connect, it is a participative, cooperative, and social Web, where users can create social activities and communicate between themselves on the network and enables users with a handful of new concepts like blogs, social media, and video-streaming platforms like Twitter, Facebook, and YouTube [40].

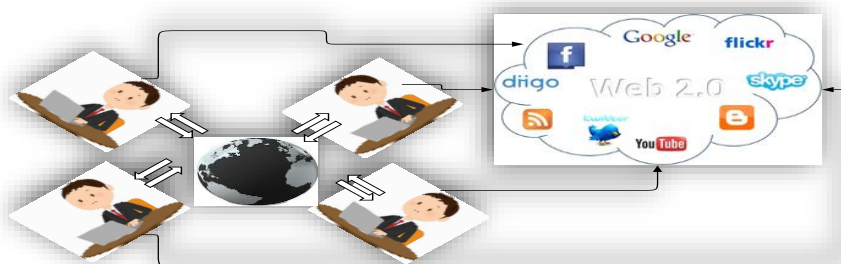


Figure 6: Web 2.0 illustration

Web 2.0 technology infrastructure contains some rules such as Atom, RSS, and RDF that are used by the designer for creating Web 2.0 services, also, Web 2.0 uses Ajax technology¹⁵ such as JavaScript, Document Object Model DOM¹⁶, REST¹⁷, XML, and CSSBut¹⁸, these properties consider issues because the user can be hacked in privacy and personal information security [39].

The increasing number of Web pages and requests urges Web applications to invent new methods for document handling, i.e., computers are becoming abided to understand the data they are processing [41]. The main idea was to provide a context to the linked documents in a machine-readable manner, i.e., to implement the transition from the Web of documents to the Web of data (see Figure 7).

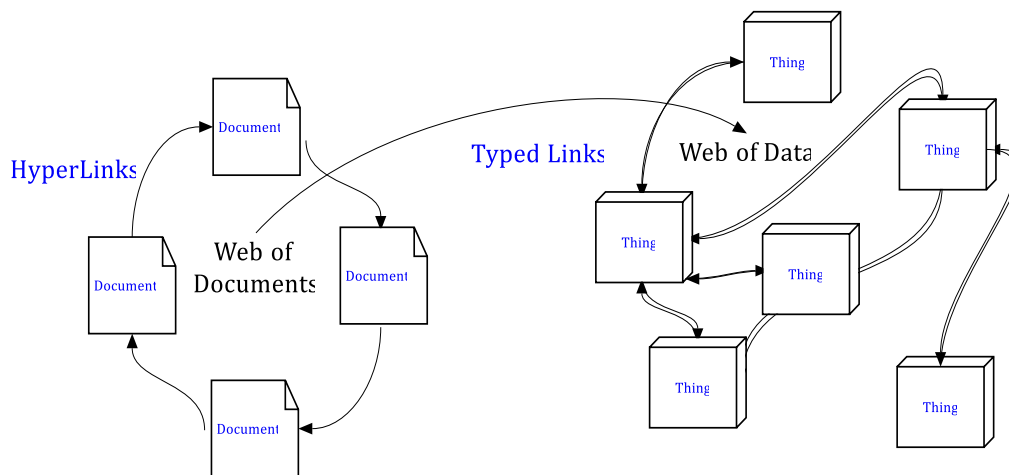


Figure 7: The transition from Web of documents to the Web of data

In 2010, **Web 3.0** (also known as *The Semantic Web, the Web of Data, or the Internet of things*) was introduced (see Figure 8) [37], Web 3.0 is an executable platform that enables users to interact with dynamic applications.



Figure 8: Web 3.0 illustration

Web 3.0 also, enables software, databases, and services of the Web to use and understand that information in a much more intelligent way. *Conrad Wolfram's* theory about Web 3.0 tried to enable computers to “*think intelligently*” for new data searches instead of humans [35]. Table 1 compares the development stages of the Web since 1995. Web 3.0, aims at modeling

¹⁵ [AJAX Technologies - javatpoint](#)







¹⁶ [What is the Document Object Model? \(w3.org\)](#)

¹⁷ [What is REST - REST API Tutorial \(restfulapi.net\)](#)

¹⁸ [CSS Tutorial \(w3schools.com\)](#)

computers to act like humans when describing the specific information at high speed and bring the information for the user as the meaning of the word and do not search for the same word on the Web, e.g., google is a Web 3.0 technology infrastructure of [39].

Table 1: Web development stages since 1995

	Web 1.0	Web 2.0	Web 3.0
Web Version	 <ul style="list-style-type: none"> • Broadcast • Web of information • OPEN Access 	 <ul style="list-style-type: none"> • Share • Web of people and social Information • OPEN Contribution 	 <ul style="list-style-type: none"> • Semantic Interaction • Semantic Web of Knowledge • OPEN for innovation
No. of Sites	250,000	80,000,000	800,000,000
No. of Users	10,000,000	>100,000,000	>2,000,000,000
Features	<ul style="list-style-type: none"> • Pushed Web • Text with Graphics • One-way communication 	<ul style="list-style-type: none"> • Two-way Web • Blogs • Video • Podcasts • Sharing, and Personal publishing 2D portals 	<ul style="list-style-type: none"> • 3D portals • Avatar representation • Interoperable profiles • MUVES • Integrated games • Education and business • media flows virtually
Technologies used	HTML, HTTP, XML, XHTML, and CSS	JavaScript, and XML, DOM, REST, XML and CSS	RDF, RDFS, OWL, and SPARQL
Main Features	<ol style="list-style-type: none"> 1. Hyper linking and bookmarking on pages. 2. No communication between user and server. 3. Static Websites. 4. allows only content browsing 	<ol style="list-style-type: none"> 1. Better interaction. 2. Includes functions like Video streaming 3. Online documents. 4. Introduction of Web applications. 5. Everything becomes online and stores on servers. 	<ol style="list-style-type: none"> 1. Smart, Web-based applications and functionalities. 2. Merging of Web technology and Knowledge Representation (KR).
System Type	Ecosystem	Participation	Self-Understanding
Associated Websites			
Active Period	1990-2000	2000-2010	2010-2021

The *Web of Data* can be visioned as an extra layer that is tightly intertwined with the classic *Web of Document* and has many of the same properties:

- *The Web of Data is generic and can encompass any data type.*
- *Entities are connected by RDF links, creating a universal data graph that traverses data sources and permits exploring new data sources.*
- *Using HTTP/RDF as a standardized data access/data model mechanism eases data access compared to Web APIs, which depend on heterogeneous data models and access interfaces.*
- *Data is self-describing by dereferencing the URIs in case of encountering data description with an unfamiliar vocabulary.*
- *Representing disagreements and contra dictionary information about an entity is permissible.*
- *Publishing to the Web of Data is open to all and Data publishers are not forced in their selection of vocabularies with which to represent data.*

To make things comprehensible, the main concept of the Semantic Web can be thought of as enabling computers to understand that for example, when we talk about the capital of Libya, the answer is precisely the city *Tripoli* in *Libya*, and it will not contradict with the city *Tripoli* in the country *Lebanon*.

2.3 Linked Data

2.3.1 Linked Data Principles and the 5-star Open Data Model

In 2006, *Tim Berners-Lee* stated that “*the Semantic Web is not only about putting data on the Web. It is about making links, so (human or not) agents can explore the Web of data*” [172]. He published a deployment scheme for open data, based on five represented as “stars” [173]. A 5-star open dataset should adhere to all of these requirements:

- * Available on the web, any format provided data has an open license;
- ** Available as machine-readable structured data (e.g., Excel instead of image scan);
- *** Available non-proprietary format (e.g., CSV instead of Excel);
- **** Make use of open standards from W3C (RDF and SPARQL) and URIs to identify things;
- ***** Link data to other providers' data to provide context.

The Linked Data principles¹⁹ introduced by *Tim Berner-Lee*, encourage interlinking and publishing structured data using Web standards. The Linked data has numerous advantages over other models [174], namely: i) IRI can be accessed by Web infrastructure and typed links between data from diverse applications; ii) RDF model allows merging and consuming from diverse sources with no need for complex transformation, and iii) explicit semantics of data expressed in OWL ontologies or RDFs which can be mapped or aligned to data models of other

¹⁹ <http://www.w3.org/DesignIssues/LinkedData.html>

applications using techniques such as ontology matching. The Linked Open Data (LOD) cloud²⁰ consisted of 12 Linked Datasets in May 2007, grew to almost 300 in 2011, and by May 2020 counted up to 1301 datasets with 16283 links and more than 200 billion linked data triples have been published in different domains, including pharmaceutical, agricultural, and industrial sectors [accessed on 20/08/2021] (see Table 2 for initial LOD cloud).

Table 2: Initial Datasets in the LOD cloud as in 2007

Datasets	Description
<i>DBpedia</i>	It is a Linked Data version of Wikipedia.
<i>Geonames</i>	Contains a Linked Data version of geographical data.
<i>DBLP</i>	A bibliographic database for computer science contains a Linked Data version of academic data.
<i>Project Guttenberg and RDF Book Mashup</i>	Contains RDF data about books.
<i>Revyu</i>	which contains reviews and rating sites for the Web of Data in the form of LD.
<i>MusicBrainz, DBtune, and Jamendo</i>	Contains RDF data about the music business.
<i>FOAF</i> (Acronym of Friend of a Friend)	An ontology containing LD that describes information about people, their relations, their activities, and, more generally, social network data.
<i>World Factbook and U.S. census data</i>	Contains governmental data in the form of RDF triples.

The datasets in the LOD are updated and maintained regularly by the Insight Center for Data Analytics²¹ cloud and are categorized and appeared in different colors based on their domains: media, geographic data, publications, user-generated content, government, cross-domain, and health & life sciences. Uploading datasets in the cloud is publicly available only if it corresponds with the LOD Cloud principles accessible at (<https://www.lod-cloud.net/>), which are a marginally different version of the Linked Data principles originally published by *Tim Berners-Lee* (above). The rating system assigns stars (1-5) for each dataset, the higher number of stars denotes the quality rating of the dataset.

The evident increase in dataset size between 2007-2020 proves the increasing interest of the LOD community cloud. There exist several statistical Linked Open Data indexes available concerning the Linked Open Data clouds such as LODLaundromat²² and LODStats²³.

Herein, we would like to point to the distinction between public data and open data. While *public data* are made freely available to the general public, they are not necessarily open, *open data*, on the contrary, have a particular license of use and distribution[161]. However, before

²⁰ Linked Data cloud, <http://lod-cloud.net/>

²¹ Insight Center for Data Analytics <https://www.insight-center.org/>

²² LODLaundromat, <http://lodlaundromat.org/>

²³ LODStats, <http://stats.lod2.eu/>

transparency or any of the other effects can happen, public data have to be disclosed in the first place [163]. Since governments collect big data from multiple sources, the more government data that is available as open data, the greater the opportunities for the stakeholders to reuse them [164]. Government data are a subset of open data and are government-related data that are made open to the public [161].

Government data are a specifically important source of open data due to its scale, variety, breadth, and status as the main source of public sector information on a wide range of subjects [165], [166]. Not all government data can be published as open data, for reasons such as national security or privacy [165], [167]. For data to be considered as open data, it must be: complete, primary, timely, available, machine-readable, non-discriminatory, non-proprietary, and free-license [168], [169].

In this thesis we have worked with open data from Arabic countries and datasets that belong to the Linked Open Data Drug part, see right side of Figure 9, including the DrugBank dataset.

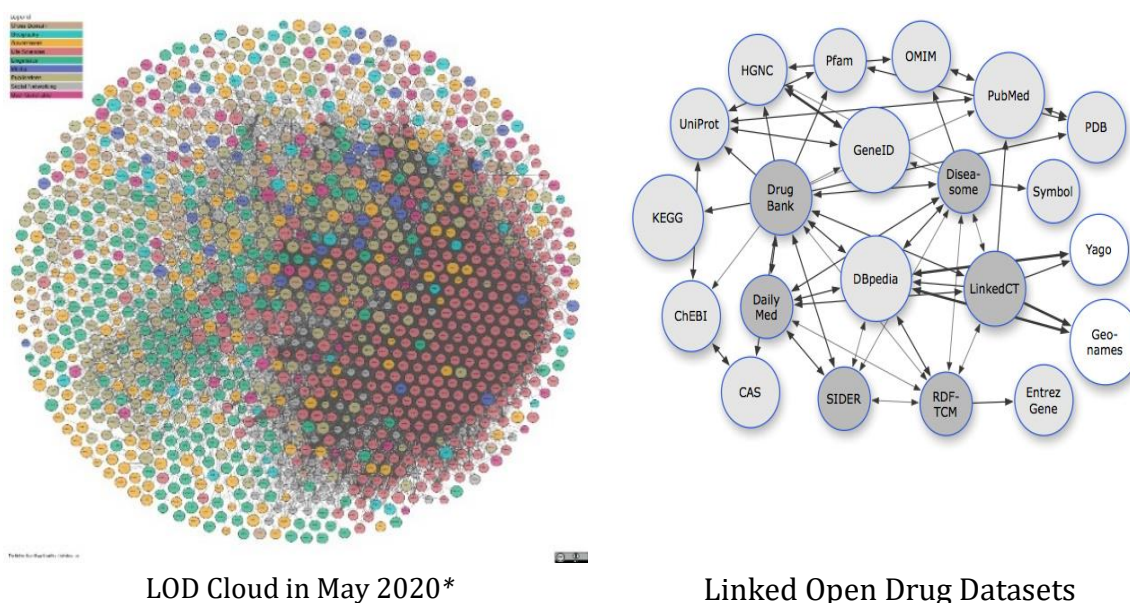


Figure 9: Linked Open Data cloud (2020) and LODD [337]

**Published regularly at <http://www.lod-cloud.net/>, and generated from the Linked Data packages described at the dataset metadata repository www.ckan.net/.*

2.3.2 Linked Open Data Best Practices

A considerable number of best practices were designed to facilitate the development and delivery of open government data as LOD but are applicable for other data too. The best practices to publish a dataset as LOD according to W3C are presented in Figure 10.

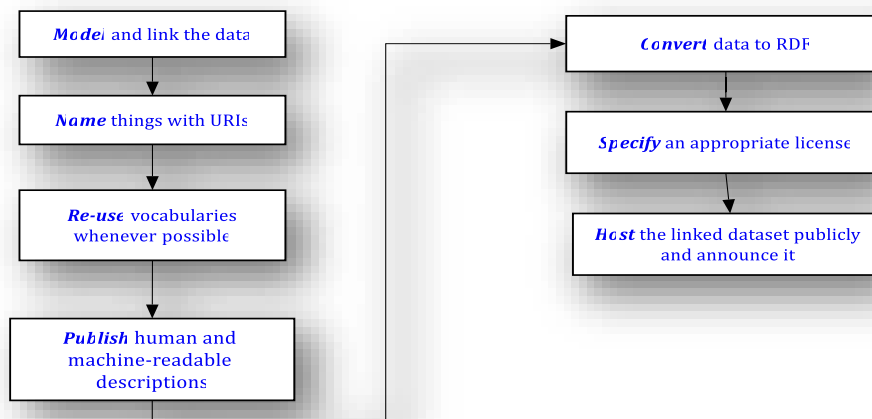


Figure 10: W3C best practices to publish a dataset as LOD

The above-mentioned principles of Linked Data have been endorsed by the Linking Open Data project²⁴, and hence, different groups from multiple domains such as the drugs industry, healthcare, media, life sciences, government, and organizations have published many interlinked datasets. Linked data best practices adaption rate increased and has led the Web to advance into a global data space comprising billions of assertions, i.e. the *Web of Data*, where both documents and data are linked[166][175]. The evolution of the Web of Data-enabled combining distributed datasets, exploring relationships between them, and supporting the development of new applications and services [121][176]. The concept of Linked Data has provided access to more data and has enabled automatic processing [175][176]. As mentioned earlier, the three key technologies which support Linked Open Data include:

- *URI (identifies entities or concepts);*
- *HTTP (a simple mechanism for retrieving resources), and;*
- *RDF (a data model for describing and linking data) [177].*

Big Data and its associates (Open data, linked data, and the LOD, etc.) are gradually developing into new scientific phenomena in many domains such as trade and industry to motivate technology to divert to data-centric architecture and operational models. There is a need to define the basic information/Semantic models, architecture components, and operational models that jointly comprise the so-called *Big Data Ecosystem*.

2.4 Semantic Web Languages

Semantic Web Languages (SWLs), such as triple languages RDF & RDFs²⁵, conceptual languages of the Ontology Web Language OWL 2 family [43] and rule languages of the RIF (Rule Interchange Format) family ²⁶, are languages used to deliver a formal description of concepts, relationships, and terms within a given knowledge domain used to write the metadata that

²⁴ <http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData>

²⁵ RDF Vocabulary Description Language 1.0: RDF Schema (w3.org)

²⁶ RIF - Semantic Web Standards (w3.org)

typically annotates any kind of Web-data. There are three families of Semantic Web languages: namely;

- Triple languages **RDF & RDFs** (*Resource Description Framework*);
- Conceptual languages of the **OWL 2** family (*Ontology Web Language*), and;
- Rule languages of the **RIF** family (*Rule Interchange Format*) [44].

As the syntactic specification is generally based on XML, the Semantics is based on logical formalisms: briefly,

- **RDFs** is a logic having intensional Semantics and the logical counterpart is *pdf* [45];
- **OWL 2** is a family of languages that relate to Description Logics (DLs);
- **RIF** relates to the Logic Programming (LP) paradigm.

Both RIF and OWL 2 have extensional Semantics.

Owning standard languages to represent and reason about domain knowledge is useless without the ability to appropriately query it. For this reason, the query language SPARQL²⁷ has been defined and considered as one of the key technologies of the Semantic Web [46].

2.4.1 RDF/XML

The *RDF/XML* is a W3C recommendation that uses XML serialization of RDF for textual representation, (<http://www.w3.org/TR/rdf-syntax-grammar/>). RDF/XML is the first standard serialization format that is based on the XML tags system. RDF triples are specified within an XML element using *rdf:RDF*, whereas, *rdf:Description* element is used to define sets of triples for a subject specified by *rdf:about* attribute. RDF/XML is recommended by OWL and SPARQL standards as their syntax input/output support for optimum competence. Most resources can be defined as an *rdf:Description* XML element with an *rdf:about* attribute that provides its URI. Several characteristics concerning a given subject are given as child elements of the corresponding XML element. *rdf:resource* attribute should be employed to refer to a given URI. Listing 1 shows the definition of one published paper using XML/RDF format. Although RDF/XML format is widely used, more human-friendly RDF sterilization appeared Include:

- 1) **RDFa**²⁸: Is a notation for embedding RDF metadata in XHTML5 Web pages, which is an extension to HTML5 that helps you markup things like People, Events, Recipes, Places, and Reviews;
- 2) **N-Triples**²⁹: This is a format for storing and transmitting data. It is an intuitive and line-based format, plain text serialization format for expressing RDF graphs on a different line, and a subset of the Turtle (Terse RDF Triple Language) format;
- 3) **N3**³⁰ (Notation 3): Are assertion and logic language that is a superset of RDF. N3 extends the RDF data model by adding variables, formulae, functional predicates logical

²⁷ SPARQL 1.1 Query Language (w3.org)

²⁸ RDFa Core 1.1 - Third Edition (w3.org)

²⁹ RDF 1.1 N-Triples (w3.org)

³⁰ Notation3 (N3): A readable RDF syntax (w3.org)

implications. It can be regarded as a textual syntax alternative to RDF/XML. It is a compact and human-readable serialization format;

- 4) **Turtle**³¹ (Terse RDF Triple Language): Is a subset of N3. It is a format to express data as an RDF data model with a syntax similar to SPARQL;
- 5) **TriG**³²: Is an extension of Turtle notation to label and represent multiple RDF graphs in the same document;
- 6) **N-Quads**³³: a superset of N-Triples, for encoding and serializing multiple RDF graphs;
- 7) **JSON-LD**³⁴: the standard JSON³⁵ based serialization format for linking data that superseded RDF/JSON format.

```

1 <?xml version="1.0"?>
2 <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
3   xmlns:dc="http://purl.org/dc/elements/1.1/"
4   xmlns:ex="http://example.org/"
5   xmlns:foaf="http://xmlns.com/foaf/0.1/">
6
7 <rdf:Description rdf:about="http://example.org/QA-ALDDA/">
8 <dc:title> Arabic Linked Drug Dataset: Consolidating and publishing
   </dc:title>
9 <dc:creator>
10 <rdf:Description foaf:name=" Guma Lakshen ">
11 <foaf:homepage rdf:resource="http://example.org/glakshen/" />
12 </rdf:Description>
13 </dc:creator>
14 </rdf:Description>
15 </rdf:RDF>

```

Listing 1: XML/RDF

However, even with the big capabilities and advantages of the RDF model, there are many challenges in using the RDF data model such as data quality assessment and validation.

2.4.2 Resource Description Framework Schema (RDFs)

In 1989, RDF Schema (abbreviated RDFs or RDF Schema) specifications were published. In 2004, the modified RDFs specifications became the W3C recommendations which follow the W3C design principles of interoperability, evolution, and decentralization [52]. RDFs is a language used to define simple resources that can be used to construct RDF statements according to the ontologies. RDFs classify resources as classes or properties. All resources of a class share the same characteristics determined by the class. Resources can be instances of

³¹ Turtle - Terse RDF Triple Language (w3.org)

³² Proposed TriG Specification (the short form) (w3.org)

³³ RDF 1.1 N-Quads (w3.org)

³⁴ JSON-LD - JSON for Linking Data (json-ld.org)

³⁵ JSON

multiple classes which in turn may have multiple instances. RDFs define the valid properties in a given RDF description, in addition to any properties or constraints of the property-type values. RDFs is a group of classes that have various properties that uses the RDF extensible knowledge which represented a model of data, the basic elements provided for ontologies description, also called RDF vocabularies. These vocabularies aim to organize the RDF resources which are saved in triple store to access by the query language SPARQL [29][53]. The RDFs contain some classes that are similar to the classes in Object Oriented Programming languages, which define the resources of class and subclass. RDFS extends RDF with Some important resources allowing for specifying well-defined relationships between classes and properties[54]:

rdfs: Resource the class of everything where all things described by RDF are resources. It can be regarded as the universal class containing everything, classes, properties, literals and even itself.

rdfs: Class is a set of things used to declare a resource as a class of other resources. It is the class of all classes including itself. In RDFs, class C can be defined by a triple in the form of:

C rdfs:type rdfs:Class

Using the predefined property and rdfs:type class rdfs:Class. Suppose, we want to utilize RDF Schema to deliver information about the categorization of medicine in Libya, the statements can be written as:

<i>Ex: Libyan-Medicine-Categorization</i>	<i>rdfs: type</i>	<i>rdfs: Class</i>
<i>Ex: Injections</i>	<i>rdfs: type</i>	<i>rdfs: Class</i>
<i>Ex: Capsules</i>	<i>rdfs: type</i>	<i>rdfs: Class</i>

rdfs: Property is a binary relation between two class individuals used to represent a property that is of type RDF property.

rdfs: subclassOf a class that has to be intended as a subset of the more general class used as a predicate, meaning that, the subject is a subclass of the object. As an example, the statement:

ex:Textbook rdfs:subclassOf ex:Book

can be understood as “That the textbook class is a subclassof the book”

rdfs: subPropertyOf declares that all things related by a given property sp1 are also necessarily related by another property sp2.

rdfs: domain used as a predicate when the subject is a property and the object is the class that is a domain of this property.

rdfs: range used as a predicate when the subject is a property and the object is the class that is a range of this property.

2.4.3 Taxonomies and Thesauri

Calaresu and Shiri (2015) defined **Taxonomy** as: "the information entities classification in a hierarchical form, according to the assumed relationships of the real-world entities that they represent" [49]. On the other hand, The ANSI/NISO³⁶ Monolingual Thesaurus Standard Defined **Thesauri** as: "a controlled vocabulary arranged in a known order and structured so that equivalence, homographic, hierarchical, and associative relationships among terms are shown and identified by standardized relationship indicators ...". Taxonomies organized and controlled vocabulary terms into a hierarchy. For example, if we consider the drugs-controlled vocabulary and say that anti-biotics is a broader term for suspension, ampule, and capsule, and that Drugs is a broader term for anti-allergics and anti-biotics, we will end up with a simple taxonomy. The "broader" relationships of taxonomy are often visually presented as a tree as in Figure 11.

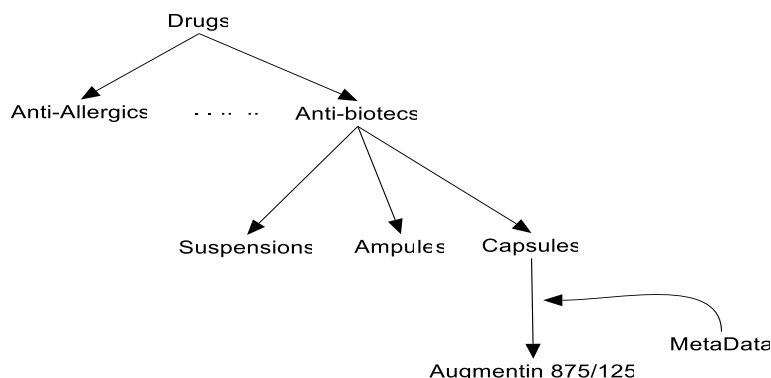


Figure 11: A simple taxonomy relationship tree

Thesaurus is regarded as a taxonomy with a set of Semantic relationships, such as equivalence, inverse, and association, that hold among the concepts. A thesaurus is used to guarantee the consistent description of concepts enabling users to refine searches and locate the required information [71]. Thesaurus stores even more metadata than a taxonomy. It might store relationship information about opposite terms (e.g., the opposite of Yes is No). Thesaurus can be used to connect a term to another term in a different vocabulary [72]. Regarding our previous example about drugs, a thesaurus might store metadata indicating that the term *augmentin 875/125* in the drug taxonomy is *Related to* the term *E. coli urinary tract infection* which is a type of urinary tract infection in a taxonomy of humans (see Figure 12).

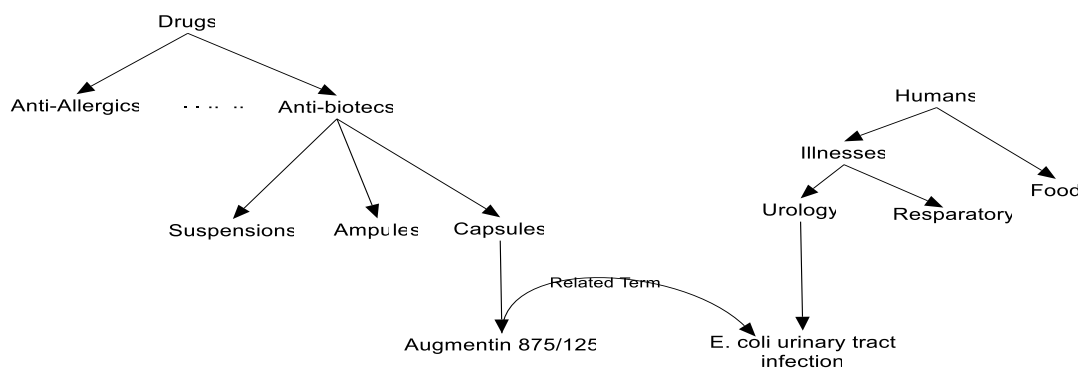


Figure 12: A drugs thesaurus relationship tree

³⁶ANSI/NISO is the American National Standards Institute/ National Information Standards Organization

Even Taxonomies and Thesauri are not designed for the Web and are not present on the Semantic Web stack, but they relate firmly to the big Semantic Web picture scenario. Both taxonomies and thesauri are utilized to improve the search user interface and strengthen the search experience.

2.4.4 Web Ontology Language (OWL)

OWL³⁷ is progressed as an RDF vocabulary extension and is derived from the Defense Advanced Research Projects Agency (DARPA)/Agent Markup Language (DAML³⁸) and Ontology Interchange Language (OIL³⁹) Web Ontology Language. The OWL and its successor OWL 2 are “*object-oriented*” languages for defining and instantiating Web ontologies. An OWL ontology can include descriptions of classes, properties, and their instances, such as:

```
class    Amoxicillin partial Penicillin-antibiotic
         restriction (hasName           someValuesFrom String)
         restriction (hasform          someValuesFrom String)
         restriction (hasUsage         someValuesFrom String)
         restriction (hasManufactureDate someValuesFrom Date)
         restriction (hasrestrictions  someValuesFrom String)
```

“The class *Amoxicillin* is a subclass of class *Penicillin-antibiotic* and has attributes: *hasName* having a string as value, *hasForm* having a value as a date, *hasUsage* a value as a string, and *hasManufactureDate* having a value as Date.

The OWL formal Semantics specifies the way to derive its logical consequences. As an example, if an individual named Zakaria is an instance of the class Student, and Student is a subclass of Person, then it can be derived that Zakaria is also an instance of Person, similarly as it happens for RDFs. However, OWL is much more expressive than RDFs, as the decision problems for OWL are in higher complexity classes than for RDFs [73].

OWL 2 is an upgraded version of OWL 1 adding several new features, including an increased expressive power [74]. IN addition, OWL 2 defines several OWL 2 *profiles*, i.e., OWL 2 language subsets tackle certain computational complexity requirements more adequately or its implementation is easier. The choice of which profile to utilize in practice depends on the ontological structure of the reasoning tasks at hand. The current version OWL 2 profiles include the following family of languages with different degrees of expressivity and computational properties [25].

- **OWL 2 Full** Informally used to denote RDF graphs, considered as OWL 2 ontologies and interpreted using the RDF-Based Semantics.

³⁷ OWL Web Ontology Language Overview (w3.org)

³⁸ DAML.org

³⁹ Cover Pages: Ontology Interchange Language (OIL)

- **OWL 2 DL** Informally used to denote OWL 2 ontologies interpreted using the Formal Semantics of Description Logic (“Direct Semantics”).
- **OWL 2 EL** is especially beneficial in applications that use ontologies that comprise huge numbers of properties and/or classes.
- **OWL 2 QL** Used for applications that massive volumes of instance data, and where query answering is the most important reasoning task.
- **OWL 2 RL** Targeting applications that necessitate scalable reasoning without compromising too much expressive power [75][76].

To model knowledge about a specific domain, OWL uses three ontology notions that are axioms, entities, and expressions explained as follows:

An axiom Referred to as a *statement* that is asserted to be true in the domain being modelled [77]. An example of a statement is “Tripoli is the capital city of Libya.” Using a subclass axiom, it can be said that class a:Capital is a subclass of class a:Country.

An Entities Represent the basic elements of the modelled domain. For example, a class a:individual used to model a group of all individuals. Similarly, the object property a:parentOf may be used to model the relationship parent-child. Also, the person a:ALI may be used to represent a particular person called "ALI".

An expression Represents complex concepts in the modelled domain. For example, a class expression prescribes a group of individuals in terms of the constraints on the individuals' attributes. Although complex language constructs allow representing more knowledge, computation becomes inefficient and eventually undecidable [78].

2.4.5 The SPARQL Query Language

Simple Protocol and RDF Query Language (SPARQL) is a declarative query language to manipulate data represented as RDF triples. SPARQL⁴⁰ Protocol and RDF Query Language⁴¹ are protocol and query languages for retrieving and manipulating RDF data. SPARQL is standardized by W3C in 2008, as a Semantic Web language used for query graph data which is represented by RDF triples. SPARQL is recommended by the Data Access Working Group (DAWG⁴²) under W3C, also it is the basic technology of the Semantic Web. SPARQL is a query language for RDF, where a query is represented by a graph pattern to match against the RDF graph. The graph patterns comprise triple patterns that resemble RDF triples, but with the option of query variables in place of RDF terms in the subject, predicate, or object positions. In the Linked Data community, it is common to see publicly accessible SPARQL endpoints where queries are sent and received over HTTP [80][29][53].

The data repositories of RDF are supporting SPARQL directly or by dedicated tools of SPARQL. Also, the SPARQL has many features computed query achieved by sub-graph

⁴⁰ <http://www.w3.org/TR/rdf-sparql-query/>

⁴¹ <http://www.w3.org/TR/vocab-data-cube>

⁴² Data Access Working Group (DAWG) (carleton.edu)

matching. SPARQL is used to express queries across local and remote data sources, whether the data resides in RDF files or databases. SPARQL tends to save development time and cost by allowing client applications to work with only the data they're interested in [81]. SPARQL Basic Syntax query to find the books authored by Tim Berners Lee is presented in listing 2.

```
PREFIX    dbr : http://dbpedia.org/resource/
PREFIX    dbo : http://dbpedia.org/ontology/

SELECT   ?book

WHERE
  {
    ?book dbo:author dbr: Tim Berner-Lee
  }
```

Listing 2: Example of SPARQL query

SPARQL builds on other standards including RDF, XML, HTTP, and WSDL, allowing reuse of existing software tooling and promoting good interoperability with other software systems. For instance, results obtained from SPARQL can be expressed in XML: XSLT to be used to generate query result displays for the Web. It's relatively easy to issue SPARQL queries, given the diversity of HTTP library support in Python, Perl, Ruby, PHP, etc. [82].

2.5 Big Data and the Web – a state of the art

There are three varieties of data types *structured, unstructured, and semi-structured* [83]. Shankaranarayan et al. (2003), proposed variant data types include *raw data items, information products, and component data items* [84]. Giant Information companies, such as Google, Yahoo, and Facebook originated this nomenclature to analyze huge amounts of data [85]. According to International Data Corporation IDC⁴³, everyone online creates an average of 1.7 megabytes of new data every second by 2020, and only 37% of all big data could be analyzed⁴⁴.

2.5.1 Big Data Definitions and Characteristics

Data in the Web is growing at a tremendous rate according to [86]–[88]; this data represents 2.5 quintillion bytes (Exabyte (EB) = 10^{18} bytes). In the year 2000, more than 800,000 Petabytes (1 PB= 10^{15} bytes) of data were stored on the Web. By the end of 2019, this volume is expected to reach 35 Zettabytes (1 ZB= 10^{21} bytes) and is also expected to grow 61% and exceed 175 zettabytes by 2025 as per International Data Corporation (IDC) expectations [89] [90]. The 3rd quarter of 2019, showed that 4.33 billion active internet users [91], which represents 8.2% growth in active internet users globally, this translates to 59% of the world population is online, and the percentage grows 8 times faster than the world population. There are today more than 1.7 billion Websites [92]. The projected global revenue from eCommerce retail in 2020 is projected to top \$4.2 trillion.

⁴³ International Data Corporation

⁴⁴ [Big Data in Digital Forensics: The challenges, impact, and solutions – MSAB](#)

Big Data Definitions

In literature, the term “*Big Data*” holds different definitions that emerged over time [38][93][94][95][96][83][97][98][99]. Big volumes of data that demand advanced techniques for capturing, preparing (cleaning), processing, storing, and analysis are called “*Big Data*”. Generally, big data refers to the datasets that couldn’t be perceived, managed, acquired, and processed by classical Information and Communications Technology ICT⁴⁵ and software/hardware tools within a tolerable time [100]. Definitions vary from one sector to another according to its utilization, for example, Big Data to Amazon or Google may vary compared to a medium-sized company, insurance broker, or telecommunications organization. Thus, definitions of big data also depend upon the industry intended for [100]–[102].

McKinsey Global Institute⁴⁶ (2011) defined big data as “*datasets whose volume is beyond the capacity of traditional database software tools to capture, store, manage, and analyze*” [86]. Gandomi and Haider (2015) reported that “*it is mostly due to fast advances in technology, exactly what can be considered big data is always changing, making it hard to express in specific and measurable terms*” [101], [103], they also reported that, “*if one dimension changes, the likelihood increases that another dimension will also change as a result*” [101]. Douglas Laney (2001) envisioned the future changes relating to the expanding size of data, through his definition of data by using a three-dimensional view as: “*Big data is high volume, high velocity, and/or high variety information assets that demand new ways of processing to enable enhancing decision making, insight discovery, and process optimization*” [83]. Loukides (2010) emphasized data volume by defining big data as “*when the size of the data itself becomes part of the problem and traditional techniques for working with data run out of steam*” [104]. Stonebraker (2012) defined big data as “*Big data can mean big volume, big velocity, or big variety*” [105]. De Mauro and Andrea et al. (2015) conducted a wide literature review on big data definitions, the review concluded that a consensual definition of big data would be: “*Big data represent the information assets featured by such a high volume, velocity, and variety to demand specific technology and analytical methods for their conversion into value*” [106]. Klievink et al. (2017) stated that “*big data is characterized by using and combining of multiple, large datasets, from various sources, external and internal to the organization*” [103], particularly in:

1. *Use of incoming data streams in real-time or near real-time;*
2. *Development and application of advanced analytics and algorithms, distributed computing, and/or advanced technology to handle very large and complex computing tasks;*
3. *Innovative use of existing datasets and/or data sources for new and different platforms, tools, and services.*

Therefore, big data analytics requires unique platforms, tools, and services that reduce time and can offer distributed and scalable solutions, such as those included in the Apache Hadoop ecosystem [107], [108].

⁴⁵ [Information and communications technology - Wikipedia](#)

⁴⁶ <https://www.mckinsey.com/>

Big Data Characteristics

Big data is usually described by its “*characteristics*” and properties sometimes called “*dimensions*” that assist to comprehend both the advantages and challenges of big data initiatives. *Laney (2011)* proposed three dimensions that characterize the challenges and opportunities of increasingly large data volumes: *volume, velocity, and variety*, later became known as the *3 Vs of big data*, see Figure 13 [109] [83].

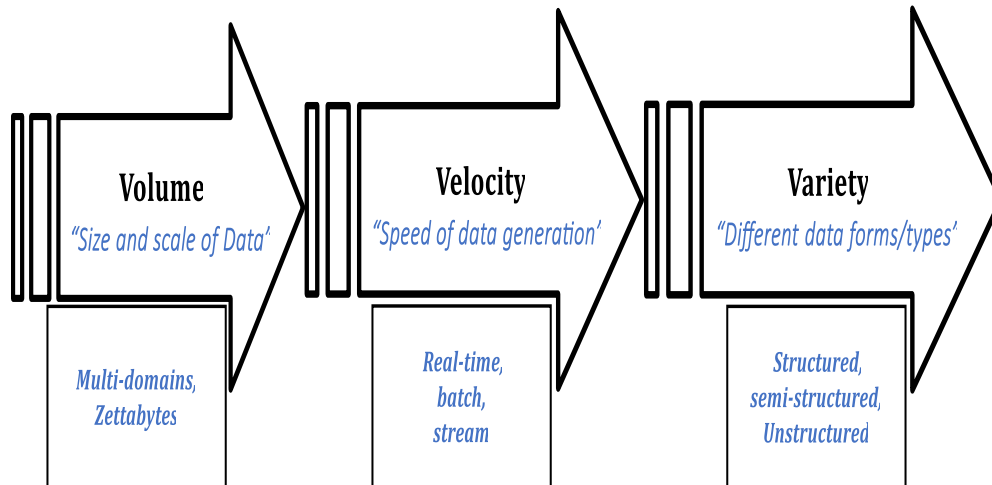


Figure 13: The original 3Vs of big data

Big data combines a set of data management challenges to work with data under new scales of size and complexity. The majority of these challenges are not new. Big data, however, is confronted with challenges raised by its characteristics related to the 3V's:

- I. **Volume** (size of data): Refers to large scales of data within data processing such as Global Supply Chains, Global Financial Analysis, Global Health issues (e.g., newly generated data due to COVID-19).
- II. **Velocity** (speed of data): Refers to streams of a high frequency of incoming real-time data (e.g., Internet of Things, Electronic Trading, Pervasive Environments, Sensors).
- III. **Variety** (data types and data sources): Refers to data using varying syntactic formats (e.g., Presentations, Spreadsheets, XML, RDF, DBMS), schemas, and meanings (e.g., Enterprise Data Integration).

These 3V's of big data challenge the traditional approaches and techniques to require new ways of data processing to empower enhanced insight discovery, decision-making, process optimization, and data visualization. As the field of big data developed, additional V's have been introduced over the years, such as the *5V's of big data* [110], but the most recognized is the list presented in Table 2 which shows the *10V's of big data* including validity, vulnerability, volatility, and visualization which sums up to the *10V's of big data* and their relation to quality [111].

Table 3: Big data characteristics and their relation to quality

	Characteristic - Description	Quality Measures (ISO/IEC 25012:2008)
3Vs	Volume - the amount of data that has to be collected, processed, stored, and displayed.	
	Velocity - the proportion at which the data is being generated, or analyzed.	
	Variety - variations in the data structure or differences in data sources themselves.	
5Vs	Veracity - truthfulness (uncertainty) of data, provenance, authenticity, accountability.	Credibility, Traceability, Provenance
	Validity - suitability of the selected dataset for a given application, accuracy, and correctness of the data for its intended use.	Accuracy, Completeness, Compliance
7Vs	Volatility - temporal validity and fluency of the data, data currency and availability, and ensures rapid retrieval of information as required.	Availability, Accessibility, Currentness, Recoverability
10Vs	Value - (useful) information extracted from the data, relevance and usefulness of data to make decisions and capacity in transforming information into action.	Efficiency, Portability
	Visualization - properly displaying and showcasing information, data representation and understandability of methods (data clustering, parallel coordinates, sunbursts, circular network diagrams, or cone trees).	Precision, Understandability
	Vulnerability - security and privacy concerns associated, weakness and other variables related to data security concerns.	Confidentiality
	Variability - the changing meaning of data, inconsistencies in the data, biases, ambiguities, and noise in data.	Consistency

It is apparent that defining big data and its characteristics will be an ongoing endeavor, but it seems that it will not negatively impact big data handling and processing. *Suthaharan* (2014) argued that the first 3V's (*volume, velocity, and verity*) cannot support early detection of big data characteristics for its classification and proposed the 3Cs as follows [110]:

- **Cardinality** defines the number of records in the dynamically growing dataset at a particular instance;
- **Continuity** defines the representation of data by continuous functions and the continuous growth of data size concerning time; and
- **Complexity** defines the large varieties of data types, high dimensional datasets, and the speed of data processing is very high

2.5.2 Importance and Benefits of Big Data

In August 2010, Ex-President Barrack Obama announced the "Transparency and Open Government" in the "Memorandum for the Heads of Executive Departments and Agencies", proclaiming that Big Data is a national challenge and priority along with healthcare and National Departments of Defence and Energy, and the Defence Advanced Research Projects Agency DARPA⁴⁷ announced a joint R&D initiative in March 2012 that will invest more than \$200 million to set up new big data techniques and tools. Its goal is to mature our "...understanding of the technologies needed to manipulate and mine massive amounts of information; apply that knowledge to other scientific fields as well as address the national goals in the areas of health, energy, defense, education, and research" [112]. The US government emphasized how big data creates "value" – within and across disciplines and domains. Value originates from the ability to analyze the data to develop actionable information⁴⁸.

Big Data is reshaping entire industries and changing human behavior and culture. It is a result of the information era and is changing how people exercise, create music, and work behavior [38]. At the technological level, there exist benefits when working with huge volumes of data, accurate data and accessibility, scalability, and integration of both structured and unstructured data. The big data benefits can be classified into three groups:

- I. **Technological:** enormous data volume, accurate and accessible, scalable, and integrable.
- II. **Financial:** decrease price, increase sales and sale leads, increase Return of Investment (RoI).
- III. **Competitiveness:** new services and products, new business models, insights in consumer behavior, more customer satisfaction, increase customer loyalty, increase sign-ups, data-driven marketing, holistic vision of the organization, and personalizing the customer experience.

The best examples of big data exist in both public and private sectors, such as advertising, music, and already mentioned massive industries (healthcare, manufacturing such as pharmaceutical industries or banking), to real-life scenarios, in hotel service or entertainment. The following lists some selected examples of Big Data use-cases:

- **Healthcare is used to map disease outbreaks, test, simulate alternative treatments**⁴⁹. Companies like Nike® uses health monitoring wearables equipment's to track customers and provide feedback on their health. Currently, during the COVID-19 Epidemic, Big Data is being utilized in many applications as Identification of infected cases, Travel history, Fever symptoms, Early-stage identification of the virus, Identification and analysis of fast-moving disease, Information during the lockdown, public movement in the affected areas, and Faster development of medical treatments.
- **NASA utilizes Big Data to discover the universe**⁵⁰.

⁴⁷ [Defense Advanced Research Projects Agency \(darpa.mil\)](https://www.darpa.mil)

⁴⁸ [FACT SHEET: Big Data Across the Federal Government | whitehouse.gov \(archives.gov\)](https://www.whitehouse.gov/archives)

⁴⁹ [How Big Data Is Being Used to Fight Infectious Disease Threats - insideBIGDATA](#)

⁵⁰ [What is NASA doing with Big Data today? | openNASA](#)

- **Music industry** substitutes trail-and-run procedures with Big Data studies⁵¹.
- **Services and utilities** are used to study customer and potential customer's attitudes and avoid blackouts and disasters⁵².
- **Cybersecurity** is used to stop cybercrime as e*mail & internet fraud and similar technology-related crimes⁵³.
- **Telecommunications** are used in customer acquisition, network optimization, and customer retention.
- **Financial services** used for customer analytics to personalize their offers, risk assessment, fraud detection, and security threat detection ...
- **Insurance** is used to help efficiently in pricing, underwriting and risk selection, management decisions, and loss control and claim management.

2.5.3 Big Value Created from Big Data

Big data can produce value from analyzing and visualizing big data for different purposes like data analysis related to diseases as COVID-19 pandemic or comprehending root reasons of a specific product revenue decline. *Brown et al.* (2011) proposed a three-step approach that can contribute to determining how to get value from Big Data [113] as follows:

- **Start with the Right Big Data Store** that is closely related to business needs, accomplished by matching the business problem or opportunity with the right technology.
- **Building domain knowledge** involves building the necessary expertise that determines which data, from all the possible sources, are valuable and which are not.
- **Choose the right reporting and analysis tool** that enables the right overall big data approach.

Choo (1996), stated that big data value can be described in the context of the dynamics of knowledge-based organizations [114]. Figure 14 depicts an overview of the big data and knowledge discovery process, whereas decision-making processes and organizational activities are dependent on the process of knowledge creation and sense-making.

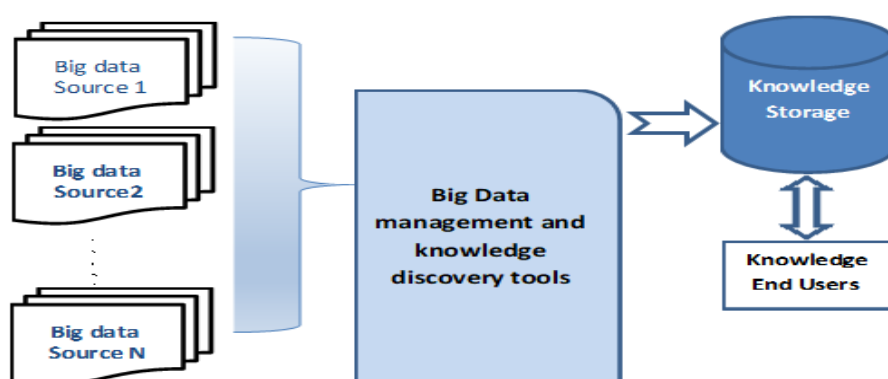


Figure 14: Big Data Knowledge Discovery

⁵¹ [Predicting the Next Big Hit - Big Data Science & the Music Industry \(simplilearn.com\)](http://simplilearn.com)

⁵² [Digital transformation and the utility of the future | Deloitte Insights](https://www.deloitte.com)

⁵³ [How Big Data Is Used to Fight Cyber Crime and Hackers: Fascinating Use Case from BT \(forbes.com\)](https://www.forbes.com)

The big data value can be described in the context of the dynamics of knowledge-based organizations [114], whereas decision-making processes and organizational activities are dependent on the process of knowledge creation and sense-making as illustrated in Figure 15. Regardless of how data is managed within any foundation, if it is processed properly, it can produce tremendous business value.

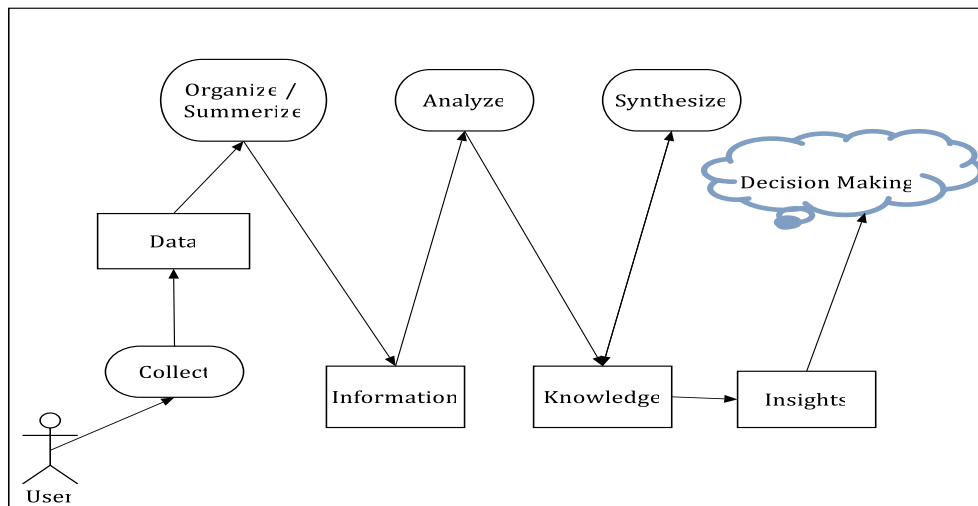


Figure 15: Big Data Value chain

Value Chains are utilized as a decision support tool to formulate the activities chain of an organization performs to deliver a valuable service or product to the market [115]. Curry (2014) identified a five-step Big Data Value Chain: *Data Acquisition; Data Analysis; Data Curation; Data Storage; and Data Usage*, that can be used to model the high-level activities that comprise an information system [116]. We identified five generic ways: *supporting experimental analysis; creating transparency; supporting real-time analysis and decision; facilitating computer-assisted innovation in products; and assisting in defining market segmentation* (illustrated in Figure 16) that big data can support value creation for organizations, where value originates from the ability to analyze the data to develop actionable information [117].

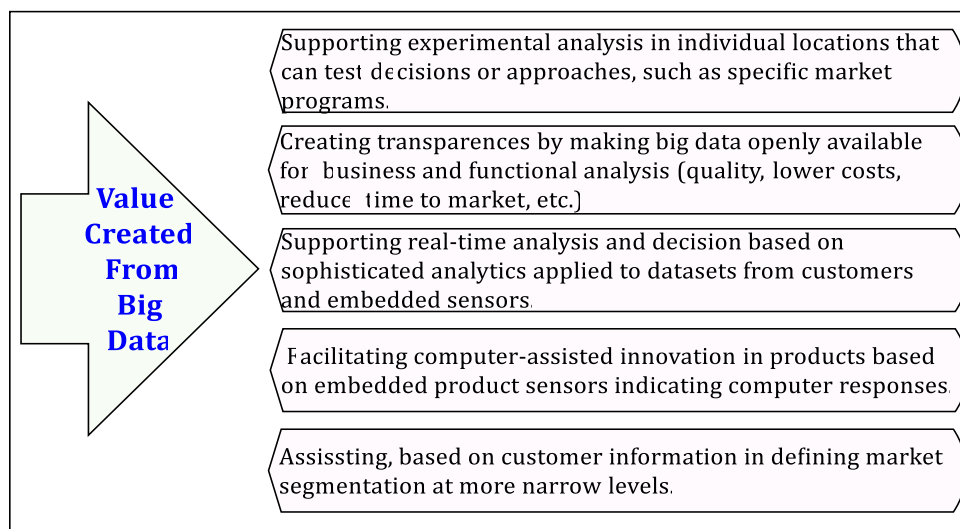


Figure 16: Five-ways supporting value creation from big data

2.5.4 Challenges of Big Data

The features of data combined with targeted business goals face many challenges while dealing with big data. Several challenges emerge in various dimensions confronting the use of big data mentioned in the literature, namely: *Data management; Data Heterogeneity; Artificial Intelligence (AI) and Machine Learning (ML); Data storage and analysis; Scalability and data visualization; Uncertainty of Data; Knowledge discovery; Data security; Human resources and manpower; and The appearance of new technologies* were studied and reviewed in [118][119][120][121][122]. Big data challenges such as *data quality, integration, governance, and manipulation* arise as potential key points that should be accounted for during constructing a Big Data management solution [123]. Other challenges were also identified by other researchers, namely: *heterogeneity and incompleteness* [124]; *high-dimensional data* [125]; *large-scale models* [125]; *failure handling* [124]; *energy management; and human resources and manpower*. It is worth mentioning that big data challenges exceed the technical levels. A crucial challenge that arises is to provide suitable data processing solutions for efficient and effective integration of data, process management, and suitable analysis tools [126]. Linked data also, faces a new set of challenges of data quality concerning a variety of aspects stated in [5][127][6]. However, Big Data analysis challenges can be identified in four groups as *i) data storage and analysis; ii) knowledge discovery and computational complexities; iii) scalability and visualization of data, and iv) information security* [128].

2.5.5 Tools and Technologies of Big Data

Big Data technologies and tools are software utilities designed for processing, analyzing, and extracting information from large data which can't be handled with traditional data processing software. Institutions, Companies, and organizations required big data processing technologies to analyze the massive amount of real-time data. They use Big Data technologies and tools to come up with predictions to reduce the risk of failure as their work progresses. The most popular big data tools are:

- **Hadoop MapReduce** a software framework for distributed processing of large volume datasets on computer clusters of commodity hardware;
- **Scala** is an object-oriented language that is mainly appropriate for pattern matching;
- **Apache Giraph** is an extension of Hadoop's MapReduce framework to execute graph processing on Big Data; and
- **Tableau** is a business intelligence tool used to generate reports, charts, graphs, and dashboards.

Big Data Storage tools have a dual purpose, they offer an infrastructure on which is possible to run analytics tools, and concurrently a place to store and query Big Data. The most relevant variables in choosing a Big Data storage tool include the *existing environment; current storage platform; growth expectations; size and type of files; database and application mix* [129]. The following tools are gradually used in big data and real-time Web applications, due to their ease of design and scalability.

- **JSON** is a universal format suitable for exchanging information between applications over numerous protocols.
- **RESTful** is an API that allows the communication between a Web-based client and server that employs representational state transfer (REST).
- **SQL/NoSQL** offers mechanisms for the storage and retrieval of data. NoSQL databases are appropriate for data that changing or evolving repeatedly [130].

Existing tools of big data can be categorized into three groups, as follows:

- I. **Computing tools:** Apache⁵⁴ Hadoop⁵⁵, Apache Spark⁵⁶, MongoDB⁵⁷, Apache MapReduce⁵⁸, Apache Pig⁵⁹, Cloudera impala⁶⁰, IMB Netezza⁶¹, Apache Giraph⁶², QlikView⁶³, QlikSense⁶⁴, Scala⁶⁵, Apache Storm⁶⁶, Presto⁶⁷, Apache Flink⁶⁸, Rapidminer⁶⁹, Knime⁷⁰, Elasticsearch⁷¹, and Tableau⁷².
- II. **Storage tools:** Apache HBase⁷³, Apache Hive⁷⁴, Apache Cassandra⁷⁵, Apache Kafka⁷⁶, Apache Sqoop⁷⁷, and Neo4j⁷⁸
- III. **Supporting technologies:** JSON, SQL and NoSQL, RESTful⁷⁹, and Machine-to-Machine⁸⁰.

A detailed article on TechVidvan Webpage (<https://techvidvan.com/tutorials/big-data-technologies/>) provided more information.

⁵⁴ [Welcome to The Apache Software Foundation!](#)

⁵⁵ [Apache Hadoop](#)

⁵⁶ [Apache Spark™ - Unified Analytics Engine for Big Data](#)

⁵⁷ [The most popular database for modern apps | MongoDB](#)

⁵⁸ [Apache Hadoop 3.3.0 – MapReduce Tutorial](#)

⁵⁹ [Welcome to Apache Pig!](#)

⁶⁰ [Apache Impala supported by Cloudera Enterprise](#)

⁶¹ [Netezza Performance Server - Overview | IBM](#)

⁶² [Giraph - Welcome to Apache Giraph!](#)

⁶³ [QlikView – Powerful Interactive Analytics & Dashboards | Qlik](#)

⁶⁴ [Qlik Sense | Data Analytics Platform](#)

⁶⁵ [The Scala Programming Language \(scala-lang.org\)](#)

⁶⁶ [Apache Storm](#)

⁶⁷ [Presto | Distributed SQL Query Engine for Big Data \(prestodb.io\)](#)

⁶⁸ [Apache Flink: Stateful Computations over Data Streams](#)

⁶⁹ [RapidMiner | Best Data Science & Machine Learning Platform](#)

⁷⁰ [KNIME | Open for Innovation](#)

⁷¹ [Get Started with Elasticsearch, Kibana, and the Elastic Stack | Elastic](#)

⁷² [We're changing the way you think about data \(tableau.com\)](#)

⁷³ [Apache HBase – Apache HBase™ Home](#)

⁷⁴ [Apache Hive TM](#)

⁷⁵ [Apache Cassandra](#)

⁷⁶ [Apache Kafka](#)

⁷⁷ [Sqoop - \(apache.org\)](#)

⁷⁸ [Neo4j Graph Platform – The Leader in Graph Databases](#)

⁷⁹ [What is REST \(restfulapi.net\)](#)

⁸⁰ [Machine-to-machine communication \(M2M\): definition and principles - IONOS](#)

2.6 Characteristics of a Modern Data Ecosystem

In ICT⁸¹ literature, an *ecosystem* is defined as “a complex network of interconnected systems”. Various data sources initiators from diverse institutions are combined and enriched in cross-industry, socio-technical networks – so-called data ecosystems [180][181]. In literature, scientists claim that involvement in ecosystems is no longer a choice, but rather a necessity for companies to unlock the benefits of data sharing [181][182][183]. *McKinsey* believes that data ecosystems will generate 30% of the world's gross domestic product by 2025⁸². However, even data ecosystems are acquiring in importance substantial number of companies are still reluctant to open their data resulting in denying the utilizing the data ecosystems capabilities [184][185]. *Davies* (2011) introduces the concept of reinforcing an Open Data Ecosystem to *assist, determine, and evaluate* possible strategies that government and non-government ODI can adopt in pursuing the achievement of the pledged benefits of open data [22]. *Harrison et al.* suggested the use of ‘*strategic ecosystems thinking*’ as a framework for recognizing where interesting problems are located in an open government ecosystem and how specific new knowledge about the coherence and interaction can report problem solutions and trigger innovation [186][187].

Characteristics of a modern data ecosystem are as follows [200]:

1. **Customer focus**, holistic operations, cross-department, and cross-product/service cooperation to incorporate the customer journey for extracting value from big data.
2. **Data-driven**, ability to gather additional information about customers, transactions, processes, which makes it possible for a global company (ecosystem) to make a better offer to its customers;
3. **Automation of processes** allows to reduce significantly the costs and prices of products and services;
4. **Globalization (globality)**, permitting the ecosystem to scale its offerings, beyond borders, region, and country;
5. **Dynamism**, implying a fast response to changes in the environment and adaptation to them, the propensity of business intelligence to make rapid decisions.

From another perspective modern data ecosystem should comply with [201]:

- **Low latency reads and updates**: This is the measurement of the system’s delay time/waiting time. The Big Data ecosystem should submit low read time and low update time as far as possible.
- **Robustness and fault-tolerance**: *Robustness* is defined as the system's ability to manage erroneous input and errors during execution. Systems require to work efficiently and properly even in machine-failure cases. The system must be sufficiently robust to cope with machine failures and human errors. *Fault tolerance* is defined as the system's ability to continue operating correctly even if some of its components fail. The systems must be human-fault tolerance.

⁸¹ [What is ICT \(Information and Communications Technology\)? \(techtargget.com\)](http://techtargget.com)

⁸² [The role of insurers in insurance ecosystems | McKinsey](http://McKinsey.com)

- **Generalization:** Big Data systems should support a wide range of applications with the operational functions of all datasets.
- **Scalability:** is the ability to preserve system performance in case of data -volume increases by adding system resources.
- **Minimal Maintenance:** defined as the work required by the system to keep it running smoothly. Big Data systems with modest implementation complexity should be prioritized, i.e., system maintenance should be kept minimal.
- **Extensibility:** When needed, the big data system provision to add functionalities with minimized development cost.
- **Debuggability:** defined as the system's ease of being debugged. When required, a big data ecosystem must provide the necessary granular information to debug and also simplify the required level to which something can be debugged.
- **Ad-hoc queries:** Big Data System should facilitate ad-hoc queries. As the need arises, ad-hoc queries can be created to obtain the required information.

2.7 Summary

In recent years, semantic-based technologies have been increasing their relevance both in the research and business worlds. W3C, together with universities and IT research organizations and in cooperation with the major software companies and government agencies, has already accepted many specifications, guidelines, protocols, software, and tools that are the basis for the realization of the Semantic Web vision. Innovative enterprises (for instance Google, Amazon, Facebook) interested in developing new business models, catching new opportunities from the Semantic Web, and offering billions of customers new services, have introduced semantic technologies to facilitate data integration and interoperability, as well as improve search and content discovery.

In this Chapter, a state-of-the-art is presented related to semantic technologies and Big Data. The analysis has been presented at TELFOR 2016 conference and has been cited 22 times.

- *Guma Lakshen, Valentina Janev, and Sanja Vraneš. 2016. "Big Data and Quality: A Literature Review". 24th Telecommunications forum TELFOR 2016. 22-23 Nov. 2016. IEEE. Belgrade, Serbia. DOI: 10.1109/TELFOR.2016.7818902.*

The notion of Data Ecosystems has been utilized by several stakeholders as well as it is reviewed in several articles. However, there is not much insight into Linked Data Ecosystem terminology. The evidence is the absence of a well-accepted definition of the term **Linked Data Ecosystem**. An adequate Data Ecosystem stakeholder's communication requires a common and unified definition of all the essential Data Ecosystem elements as well as it requires a formal definition for Data Ecosystems terminology.

Therefore, in this thesis, a Modern Data Ecosystem is viewed as a complex set of numerous interconnected components related to big data, models, and organizational structures and roles covering the whole data lifecycle [178]. Additionally, in the next

chapters, we will use the notion of **Big and Linked Data Ecosystem**, having in mind that Big and Linked ecosystems that are expanding daily, for instance, Google®, Facebook®, Twitter®, LinkedIn®, Alibaba®, and Amazon®, etc.

Chapter Three

Quality Issues of Linked Data Ecosystems

CHAPTER THREE – QUALITY ISSUES OF LINKED DATA ECOSYSTEMS

Organizations are increasingly depending on data analysis to gain data value and achieve a competitive advantage. As data size gets bigger, creating a real value from such big data is possible if data passes quality assessment tests. Fulfillment of dimensions such as *accuracy, completeness, consistency, relevancy, and reliability of data* is essential to make good decisions and actions [131]. To guarantee that data conforms with an acceptable level of quality, methods and techniques performing data quality assessment are obligatory to support the identification of suitable data to process [132]. The International Standard Organization ISO 9000, defined “*quality*” as the “*degree to which the consumer's needs are satisfied, by representing all the characteristics of the product or service requested by the customer*” [133]. The requirement is that data should be free of data quality problems and must include the “*necessary*” or “*desirable*” properties [134][135]. *Wang and Strong* (2013) and *Miller* (2015), defined data quality as “*data that are fit for use by data consumers*” [135][136]. Similar definitions exist in [54][135].

3.1 Introduction

Data quality requires new algorithms to deal with novel requirements related to variety, volume, velocity, and other issues that were not required for the traditional databases [137]. Data quality is closely linked to the technologies and processes for identifying, understanding, and correcting defects in data that support efficient information governance across decision-making and operational business processes [138][139]. When dealing with data quality, several issues need to be considered such as *errors and inconsistencies; data entry misspellings; missing and/or incomplete information, or other invalid data* [140]. A data quality assessment metric, indicator, or measure is an action for measuring a data quality dimension [141]. Data quality is assessed by employing different dimensions, which definition is mainly depending on the context of use [131]. Table 4 shows the research issues and application domains discussed in data quality literature and EU research projects [117]. In literature, linked data quality is discussed by offering several contributions proposing assessment algorithms for these consolidated dimensions, but utilizing big data generates new challenges related to their main characteristics volume, velocity, and variety [132][142][143][144]. During the transformation from unstructured “*Data lakes of information*” to integrated structured linked enterprise data, priority is given to data integration [145]. Therefore, data quality assessment becomes a secondary concern during the initial stages. *Josko and Ferreira* (2016), stated that “*data quality assessment outcomes are essential to ensure useful analytical processes results*” [146]. For linked datasets, the earlier quality is assessed, the better, as the cost of fixing a bug rises exponentially when a task progresses [147].

Table 4: Research issues and application domains in data quality literature and EU research projects

Project	Research	Applications / Case Studies
https://www.big-data-europe.eu	<ul style="list-style-type: none"> • Storage (<u>Hive</u>, <u>Cassandra</u>). • Message passing (<u>Kafka</u>, <u>Flume</u>) • Multi-purpose data processing and analysis (<u>Apache Hadoop</u>, <u>Apache Spark</u>, <u>Apache Flink</u>) • Publishing (<u>Geotriples</u>) 	<ul style="list-style-type: none"> • Test generic infrastructures are found in the health domain. • Drug discovery. • System monitoring in wind energy production unit. • Viticulture. • Crowd-sourcing in transport. • aggregation platform in the transport sector. • Climate pilot.
http://byte-project.eu/	<ul style="list-style-type: none"> • Setting the stage on big data. • Elements of social impact. • Case studies in positive and negative externalities. • Evaluating and addressing positive and negative externalities. • Foresight analysis. • Roadmapping. • The big data community. • Stakeholder engagement. • Dissemination. • Project management. 	<ul style="list-style-type: none"> • Environment case study: Earth and space observation portals and associated initiatives. • Utilities / smart cities case study: Utilities and smart cities big data utilizers (various). • Cultural data case study: A Pan-European Cultural Heritage Organization (PECHO). • Energy case study: big data explorers and producers for oil and gas (various). • Health case study: A Genetic Research Initiative (GRI). • Transport case study: Shipping industry stakeholders. • Crisis informatics case study: A Research Institute for Crisis Computing (RICC).
http://optique-project.eu/	<ul style="list-style-type: none"> • Real-time stream processing. • Scalable query rewriting. • Query evaluation with Elastic Clouds. • End-user-oriented Query interface. 	<ul style="list-style-type: none"> • Health care.

One of the important elements when linking multiple data sources is meeting up with data quality requirements such as: *accessing data accuracy and consistency, consolidating different data representations, and eliminating duplicate information*. Although the LOD cloud increasingly added data published as Linked Datasets, its dataset’s quality varies considerably, ranging from expensively formatted datasets to fairly low-quality linked datasets [5]. The process of Linked Data generation normally comprises data transformation from original data sources, mapping data to several ontologies and vocabularies, and data fusion and interlinking from diverse data sources. These phases are a source of possible data quality issues in several facets [148][149][5], such as *dereferenciability of resource IRI, semantic accuracy of vocabularies, consistency, completeness, and relevancy*. In literature, several authors reviewed features such as: data quality [150][102], data integration [100], data analysis [100][151][152],

knowledge discovery [151][152][153], data visualization [128], data storage [128], and scalability [154]. These reviews broadly examine the various concepts and phases of Big Data management concentrating mostly on big data dimensions.

In this Chapter we will raise the quality issues of linked open data in modern ecosystems via the following sections:

Section 3.2 introduces a generic use-case and discusses the main components of a big and Linked Data Ecosystem, focusing on end-users, challenges, and building blocks. It elaborates an example of a Modern data ecosystem that will be used for the design of the Arabic Linked Drug Data Application (ALDDA).

Section 3.3 presents the state-of-the-art issues related to the quality of Linked Open Data, by giving diverse definitions, life-cycles from various contributors. Next, the data quality issues are discussed and how data quality problems are classified.

Section 3.4 discusses existing Strategies and Techniques of Linked Data Quality Assessment and presents a comparison of tools for Linked Data Quality Assessment. Based on the analysis, the functionalities of ALDDA-QA were developed, see Section 5.

Section 3.5 proposes **A Conceptual Methodology for Linked Open Data Ecosystems Quality Assessment** that has been used in the thesis and for developing the Arabic Linked Data Drug knowledge graph.

3.2 Generic use-case in a Linked Data Ecosystem

The main function of a data ecosystem is to capture data and produce useful insights and value. The data ecosystem deals with the evolving of data, models, and supporting infrastructure during the whole big data life-cycle which includes data collecting, storing, processing, visualizing, and delivering results to the intended users via applications [139]. Data life-cycle offers a high-level overview of the phases implicated in the successful management and maintaining of data for any use/reuse process. Particularly, multiple versions of data life cycles exist with differences attributable to variation in practices across domains or communities [144][161][189][190]. In the literature, there exist many frameworks that manage big data ecosystems [191][178][192][193][19]. Figure 17 illustrates a generic use-case of the linked data ecosystem is provided in which identifies the main participants and the use-case of the ecosystem.

A brief description of the process of Linked data ecosystem use-case is as follows:

- **The dataset owner** is the creator of the dataset and could be a public entity (e.g., a governmental organization, university, hospital, etc.), a private entity (e.g., a pharmaceutical company, media group, sports cooperation), and/or an individual owns and manages a dataset in terms of creating, storing, updating, and publishing the dataset.

- **The dataset user** the body who is corresponding to entities, applications, services that utilize the datasets for diverse reasons. The user may provide services such as selecting, discovering, querying, and analysis of a particular dataset.
- **The dataset integrator** is responsible for integrating several datasets and provides integrated access services of the combined datasets.
- **The dataset expert** is regarded as a highly qualified aggregator responsible for selecting the suitable datasets or sub-datasets, needed for additional services such as data analysis. If requested, the expert may publish the results of the analysis as a new dataset.

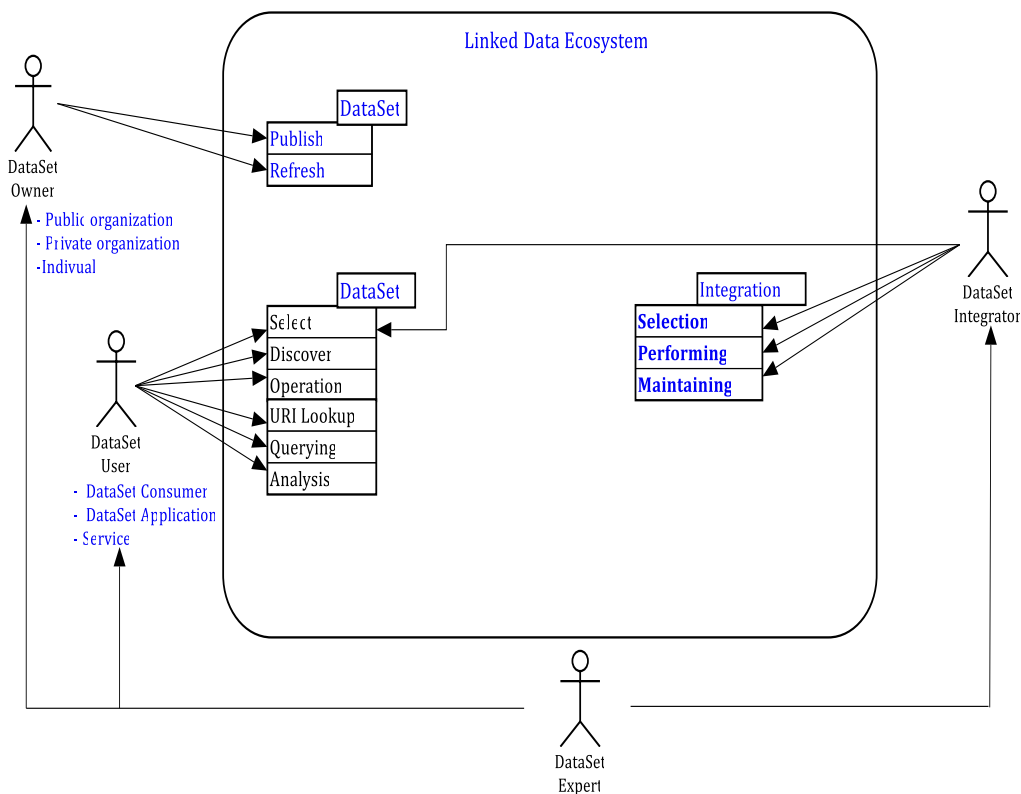


Figure 17: Generic Linked data ecosystem use-case

3.2.1 Challenges

After establishing an Open Data Ecosystem, challenges could arise placing the development or even endangering the ecosystem existing. According to *Cai and Zhu (2015)* [121], data quality is faced with the following challenges:

- **data sources diversity** results in a diversity of data types, complex data structures and increases the difficulty of data integration;
- **enormous data volume** means the difficulty to assess **data quality** within a reasonable amount of time;
- **data change very fast** and the “timeliness” of data is very short which demands higher requirements for processing technology; and

- **no unified and approved data quality standards** have been formed to meet the ISO 8000 data quality standards.

If these challenges are predicted before their existence and mitigation measures are introduced, this would reduce the risks considerably. In addition to the above challenges extra challenges exists as:

- **Privacy:** Data protection is a concern whenever data is being handled and shared, and especially when it is being published in the public domain. A clear privacy guideline for publishing Open Data should be provided.
- **Governance:** How an ODI is governed will impact its ability to achieve its objectives. Establishing a clear oversight and management structure, with precise roles and responsibilities is required.
- **Operational change:** Governmental bodies have firm internal processes which incorporate their management of data. To ensure good data quality, datasets publication as Open Data, evaluating data collection, handling, and processing should be performed.
- **Usage:** Determining how open data is being used is not directly obtainable, which is assisting when evaluating ODI, measuring its impact, and improving data provision. Data usage tracking mechanisms could be employed, such as the embedding of web-tracking code, or manual form-filling may be used.

3.2.2 Components of an Arabic Linked Open Drug Data Ecosystem

Our proposal for a data ecosystem that integrates Arabic Linked Open Drug Data is composed of three layers, see Figure 18 [196]: (1) the Data sources layer, (2) Data management and Semantic processing layer, and (3) Artificial intelligence technologies and Business Intelligence layer.

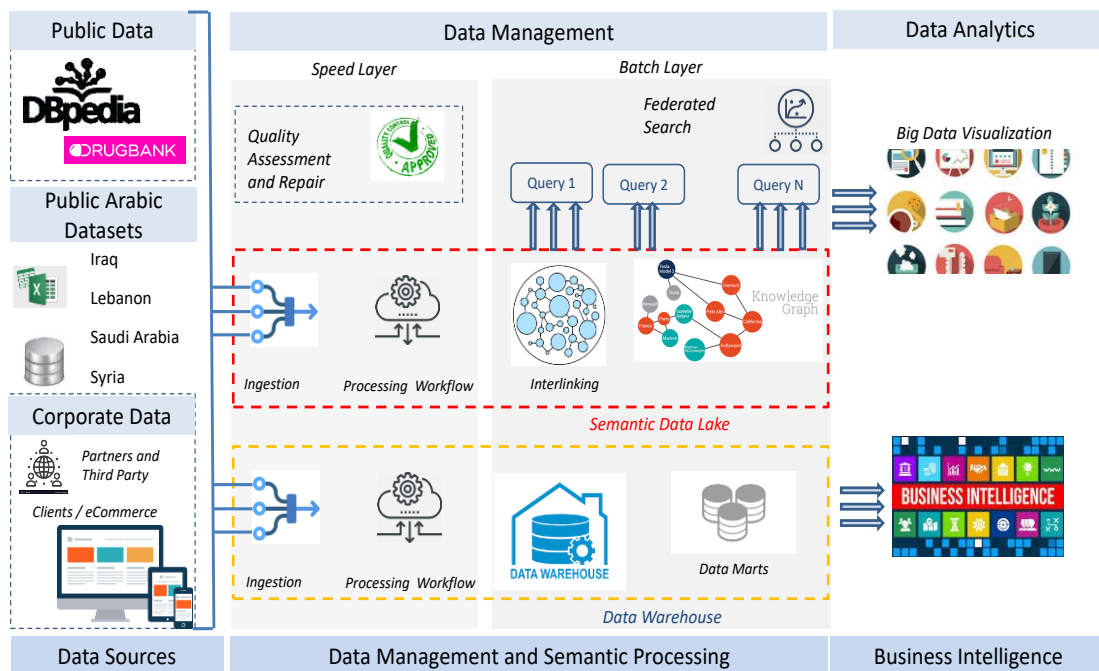


Figure 18: Modern data ecosystem for Arabic Linked Open Drug Data

1-The data sources layer is composed of both private and public data sources where dissimilar data sources and systems generate data. The interconnected systems are the property of the organization or its partners, or the data is open on the Web.

In well-organized data architecture, diverse types of data can be easily obtained and stored; on the other hand, controlling diverse datasets from different service providers is the most defiant mission. Allowing developers to create applications utilizing the open datasets, machine-readable formats are needed. There exist many open data sources from diverse domains such as; *World Bank Open Data*⁸³, *Facebook Graph API*⁸⁴, *World Health Organization (WHO) — Open data repository*⁸⁵, *Google Public Data Explorer*⁸⁶, *European Union Open Data Portal*⁸⁷, *DBpedia*, *UNICEF Dataset*⁸⁸, etc.

In this layer, different data sources and systems generate data. The interconnected systems in this layer are the property of the organization or its partners, or the data is freely available on the Web. To enable developers to create new applications on top of open datasets, machine-readable formats are needed. Languages such as, XML and JSON have quickly proven to be the predominant format for the Web and mobile applications because of their ease of integration into browser technologies and server technologies that support JavaScript. Once the data has been accumulated, the interlinking process of diverse data sources is challenging and intricate, the process is even harder if the acquired data is unstructured.

2. **Data management and Semantic processing layer**, where the data is gained through customized interfaces or crawled from the Web and conveyed using interconnected networks into storage data centers. The emerging challenges in the design of end-to-end data processing pipelines were discussed in the scientific literature as follows:

- **Different NoSQL**⁸⁹ Stores exist that lack true transactions to the time-honored SQL principles of Atomicity, Consistency, Isolation, Durability (ACID) which is a notion in database management systems that declares a set of standard properties used to ensure the reliability of a given database. Many NoSQL stores conciliate consistency in favor of availability and partition tolerance (“CAP theorem⁹⁰”) and most NoSQL stores lack true ACID transactions.

- **Data Lake** is a vast pool of raw data, the purpose for which is not yet defined. It is a notion as a new storage architecture was reinforced where raw data can be stored irrespective of source, structure, and (usually) size.

The concept of data lake was presented in the last decade to address issues connected to processing big data [197]. Moreover, the *Semantic data lakes* are presented as an

⁸³ <https://data.worldbank.org/>

⁸⁴ <https://developers.facebook.com/docs/graph-api>.

⁸⁵ <https://www.who.int/gho/database/en/>

⁸⁶ <https://www.google.com/publicdata/directory>

⁸⁷ <http://open-data.europa.eu/en/data/>

⁸⁸ <https://data.unicef.org/>

⁸⁹ [What is NoSQL? | Nonrelational Databases, Flexible Schema Data Models | AWS \(amazon.com\)](#)

⁹⁰ [CAP theorem: What is behind Brewer’s theorem? - IONOS](#)

extension of the data lake supplying it with a Semantic middleware, which permits uniform access to innovative heterogeneous data sources [198].

- **Data warehousing** is a reservoir for structured and filtered data already been processed for a precise purpose approach (based on a repository of structured, filtered data already been processed for a specific purpose) is thus envisaged as outdated as it generates certain issues concerning data integration and new data sources. Users are often confused in distinguishing between data lake and data warehousing but they are more different than they are alike. The real similarity amongst them is the high-level purpose of storing data. Some key differences between a data lake and a data warehouse are given in Table 5.

Table 5: Data Lake vs Data Warehouse

<i>Category</i>	<i>Data Lake</i>	<i>Data Warehouse</i>
Data Structure	Raw, structured, semi-structured, and unstructured	Structured and processed
Process	ELT (Extract Load Transform)	ETL (Extract Transform Load)
Purpose of Use	Not yet determined	Currently in use
Users	Data scientists, in-depth users	Business professionals, Operational users
Schema Definition	After data storage	Before data storage
Security	Highly secure	Developing
Accessibility	Highly accessible and quick to update	More complicated and costly to make changes

- **Cloud computing** emerged as a paradigm that focuses on sharing data and computations over a scalable network of nodes including end-user computers, data centers, and Web services [21]. Data pre-processing activities like data integration, enrichment, transformation, reduction, and cleansing occur in this stage, and the data is either stored in one cluster or distributed among several.

- Artificial intelligence technologies 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 to visually examine processed data. The development of business intelligence services is straightforward when all data sources gather information based on unified file formats and uploaded it to a data warehouse. However, the development of a distributed software system necessitates the interaction of services and the use of resources from varied organizations throughout the Web [199].

3.3 Quality of Linked Open Data – State of the Art

In literature, quality problems are widely recognized and confirmed by a number of studies as in [202][155][203][204][205][206][207]. According to *Google Scholar*, the number of studies on open data quality published in 2003-2014 is 4.6 times fewer than in 2018 alone. The research results show a sharp increase in the popularity of open data quality since 2017, as the number of open datasets and open data portals have started to increase. The quality of data is far from perfect, due to the effects of big data characteristics [208][121][134][209]. There is a general agreement among data stakeholders that data quality always depends on the quality of the data source [150]. Definitions of data quality are inconsistent and relate to each specific domain or context [210][3][211]. Data quality is usually defined as "*fitness for use*", meaning achieving the data quality standards that meet the users' requirements [212].

The ISO/IEC 25012:2008 Standard summarized data quality as "*the capability of data to satisfy stated and implied needs when used under specified conditions*" [137]. *Lee et al.* (2006) concluded the main causes of data quality problems as *multiple data sources, subjective judgments during data generation, insufficient computational resources balance of security and accessibility, complex data representation cross-disciplinary encoding of data, data volume, input rules that are overly restrictive or ignored, distributed heterogeneous systems, and evolving data demands* [213]. *Amadeo et al.* (1993) added problems such as data duplication, data leakage, and time calibration of multiple data sources were reported in the studies [214]. Data Quality is a main key challenge in Linked Open Data as the data is frequently transformed from multiple heterogeneous sources, semi-structured and unstructured data, which are of varying quality [215][5]. Data quality is mostly determined by weighting its features against the user's requirement. Data quality is often defined as a multidimensional perception, where each dimension is correlated to a specific user-focused aspect of quality such as accuracy, completeness, consistency, timeliness, relevancy, and accessibility [216].

Linked data quality assessment is a procedure for evaluating if data matches the user's specific needs [5]. It is usually performed using a data quality assessment framework. The dimensions are indirectly measured using one or more quality metrics. These metrics feedback values (usually between 0 and 1) can be compared to desired thresholds for pass/fail quality assessment. The assessment frameworks are capable of generating problem reports for the deformed or mislaid data they sense during metrics calculation. Data quality improvements are implemented via data corrections as required. The user performing data correction must have a clear comprehension of quality problems (faults) that exist in the data, fixing process, error locations in the dataset, which metrics are impacted by each fault, and how much improvement each fix would bring.

3.3.1 Data Quality life-cycles

Reviewing linked open data life cycle models clarified that they follow five common processes as *Data selection; Data preparation; Data publishing; Data interlinking; Data discovery; and Data reuse* [161][217]. In Literature, the data quality life-cycle generally involves

four interconnected phases: *Quality dimensions identification and related metrics; Quality assessment; Quality analysis; and Quality improvement* (see Figure 19). These phases should be adopted by an organization, the users, and the developers [189][218][219].

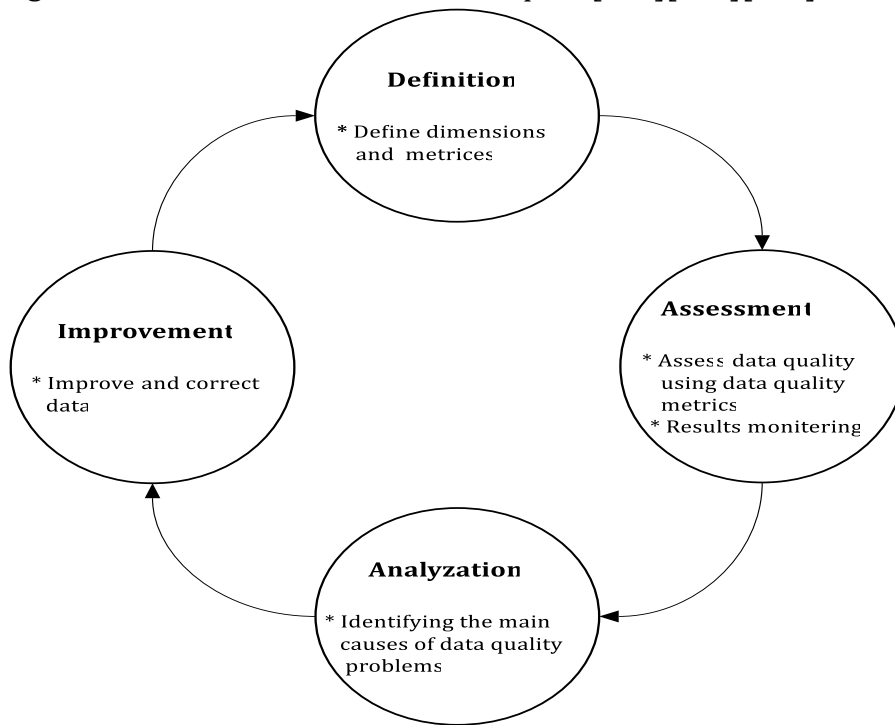


Figure 19: Generic data quality life-cycle

Most of the existing research and proposed practical solutions on data quality falls in these categories: i) *Definitions of data quality dimensions*; ii) *Open data portals and/or Open Government Data (OGD) quality assessment*; iii) *Data quality assessment frameworks*; iv) *Linked data quality assessment*, and v) *Guidelines of data quality* [220][221][222][9], where

1. The definition phase identifies the related data quality dimensions to the required context.
2. The assessment phase defines and produces metrics and assesses necessary indicators to evaluate the quality of the data.
3. The analysis phase identifies and declares the stem causes of data quality problems and calculates the impact of poor-quality information.
4. The improvement phase proposes appropriate techniques to improve data quality.

Batini et al. (2009), presented Total Data Quality Management (TDQM) to overview existing data quality assessment methodologies [223]. According to TDQM, data quality contains numerous dimensions for data users, and utilizing data quality dimensions for assessment is extensively used and existing works emphasizes on data quality dimensions and their application to datasets[155][203][224][225]. TDQM deals with data as information products and provides a comprehensive set of associated dimensions and improvements, which apply to diverse contexts. The goal of TDQM is an ongoing enhancement of the quality of information products via a cycle of defining, measuring, analyzing, and improving data and their management process, without adequate steps specified in the assessment process. TDQM and Total Information Quality Management (TIQM) are considered as the key theories for evaluating data

quality in the data quality domain through the use of data quality dimensions [226][227][135]. TDQM and TIQM follow a similar goal of offering a methodology for improving the data quality continuously. TIQM is strongly impacted by practical experience, TDQM on the other hand, emerged as a result of many years of research and suggests recognizing and documenting information production processes and information product characteristics [155][203][135].

Ferney et al. (2017) suggested analyzing datasets according to traceability, completeness, and compliance, by using a software called RapidMiner that allows for data mining [203]. *Färber et al. (2016)* analyzed the quality of openly available knowledge graphs such as Freebase, DBpedia, Wikidata, OpenCyc, and YAGO. The authors concluded that knowledge graphs have not been subject to an in-depth comparison yet, thus they provide data quality criteria to be analyzed [155]. *Wang and Strong (1996)*, *Bizer (2007)* and *Zaveri et al. (2016)* defined dimensions that authors use and apply to the above-mentioned knowledge graphs [225][228][229]. *Debattista et al. (2016)* presented a data quality life-cycle that encompasses all phases starting from data assessment, data cleaning, and data storing, the authors demonstrated that the lifecycle of quality assessment and improvement of Linked Data is an ongoing and continuous process [218].

Neumaier (2015) and *Umbrich et al. (2014)* presented an overview of automated quality assessment frameworks that allows measuring and discovering heterogeneity and quality issues in open data portals [220][230]. The authors specified and measured quality and heterogeneity issues in data portals, also, developed an automated quality assessment framework. They detected various quality issues in open data as a result of monitoring and assessing the quality of three data portals. In addition, they defined six quality metrics: retrievability, usage, completeness, accuracy, openness, and contactability. The authors presented an automated assessment framework that periodically monitors the content of CKAN portal and calculates a set of quality metrics to gain value about the evolution of the (meta-) data. The authors also suggest a common mapping for metadata occurring on analyzed portals software frameworks to improve the comparability and interoperability of portals running these different software frameworks.

These studies are beneficial in verifying Linked (Open) Data quality, but inadequate for checking autonomous dataset quality as the mentioned approaches imply profound knowledge. Diverse strategies used by the methodologies detect data quality problems, e.g., crowdsourcing mechanisms, and manual approaches. Crowdsourcing mechanisms are highly suitable for projects dealing with large to huge numbers of small tasks that require human judgment.

3.3.2 Data Quality Issues

Data quality issues can emerge at any stage of the data life-cycle, starting from the design stage of the underlying application and database, data utilization in practice, until the data extraction stage. Data quality can also influence the data at various levels to arise different root causes. Incomplete or missing data, inaccuracy, and inconsistent data are confirmed as key issues of data quality issues [231][232][233][234]. Both *Mans et al. (2015)* and *Bose et al. (2013)*

identified four broad data quality issues that could exist in process-mining event logs: *missing, incorrect, imprecise, and irrelevant data* [235][236], these four dimensions were detailed further in 27 types of data quality issues relating to the case, attribute, and event levels of the data in an event log. The widely cited Process Mining Manifesto⁹¹ suggested a 1- 5 star rating system for data quality [21]. Data quality issues are still mostly unresolved mainly due to the popularity of (open) data [237][108][238]. The most common research issues discussed in the literature are *models, techniques, tools, frameworks, and methodologies* in addition to *dimensions*, which are briefly described below:

- **Models:** *mainly used in databases to represent data and data schemas, as well as in information systems to represent business processes;*
- **Techniques:** *refer to algorithms, heuristics, knowledge-based procedures, and learning processes that help identify and solve a data quality-related problem;*
- **Methodologies:** *provide guidelines for choosing appropriate techniques and tools for effective data quality measurement and improvement;*
- **Tools and frameworks:** *software components designed, automated, and provided with an interface to evaluate the data quality assessment activities.*

It becomes visible that more errors accumulated result in more resources are necessary to reform them. *Redman and Godfrey* (1996) categorized data quality issues [224] as:

- *Issues associated with “data views” (data capturing models in the real world), as relevancy, granularity, and levels of detail.*
- *Issues associated with “data values”, as currency consistency, accuracy, relevancy, and completeness.*
- *Issues related to “data presentation”, as for appropriateness and ease of interpretation, etc.*
- *Issues related to “data safety”, as privacy, security, and ownership.*

Performing data quality functions manually is infeasible due to its tediousness, error-prone, and time-consuming character, when compared to automatic means, especially as current trends indicate that the volume of data is increasing at staggering rates [86] [239]. Also, as the data size grows, not only will data quality automation be a necessity, but new methods of automated techniques for data quality improvement and assessment will be needed. *Scannapieco et al.* (2002) performed an analysis of more than 70 existing solutions uncovering that the majority are based on definition, the grouping of data quality dimensions, and their application to datasets, which are frequently identified by researchers as problematic, even for data quality professionals [208]. *Jeczek* (2018), *Chen et al.* (2016), and *Colpaert et al.* (2013) stated that *“open data quality affects knowledge quality, reliability, and value gained from processing the data”*.

It becomes clearer, that Low-quality data may minimize the efficacy of work dramatically, therefore, greater care and attention are necessary and must be adhered to before data is added or processed. If erroneous data exists and is detected, it must be corrected or deleted before its

⁹¹ [Process Mining Manifesto - IEEE Task Force on Process Mining \(tf-pm.org\)](https://tf-pm.org/)

utilization [240]–[242]. Low-quality data also affects business decision-making, whilst high-quality data improves the efficiency of data warehousing, as data cleaning, retrieval, and downloading typically take up to 80% of the time. The “80-20” rule is a broadly utilized measurement rule, applicable to data quality, whereby 80% of the time is spent on data preparation (gathering, cleansing, and organizing), leaving only 20% to perform use and analysis⁹². Additionally, according to TDQM, the “1-10-100” rule is valid for data quality: one dollar spent on prevention will save \$10 on appraisal and \$100 on failure costs [243]. Current data quality analysis solutions are largely focused on the informal definition of data quality and measurement of acquired values, but mechanisms for determining data quality characteristics in formalized languages are less known. Similarly, there are no well-known solutions that allow users to simply analyze the quality of specific datasets by defining specific data quality requirements for individual parameters of interest [206].

Regarding Linked Data, many authors have pointed out issues such as the accuracy, *completeness, conciseness, and consistency* of open data. *Kontostas et al.* (2014) 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 [204]. At the same time, *Zaveri et al.* conducted a user-driven quality evaluation which stated that DBpedia indeed has quality problems (e.g., around 12% of the evaluated triples had issues) [277]. They can be summarized as *incorrect or missing values, incorrect data types, and incorrect links*. Based on the survey, and developed a comprehensive quality assessment framework based on 18 quality dimensions and 69 metrics. Based on the work of *Zaveri et al.*, the ISO/IEC 25012:2008 DQ model, and *Radulović et al.*, developed a linked data quality model and tested the model with DBpedia with a special focus on accessibility quality characteristics [229][21][278] respectively.

3.3.3 Data Quality Problems Classification

Singh and Singh (2010) provided a descriptive classification of data quality problems caused in data warehousing, a comprehensive list is submitted in [244]. Data quality problems are divided into two parts single-source and multi-source drawbacks [140]. As a result, the goal of classifying information quality drawback is illustrating non-standard information and distinctive actual application of knowledge for corresponding necessities. *Rahm and Do (2000)* classify data quality problems into single and multi-source problems as presented in Table 6. The classification indicates some typical data quality problems for the various cases but does not show the single-source problems that occur most likely in the multi-source case besides specific multi-source problems [245].

Schema-level problems are also reflected in the instance level; they can be addressed at the schema level by an improved schema design (schema evolution), schema translation, and schema integration. Whereas, Instance-level problems, point to errors and inconsistencies in the actual data contents which are not shown at the schema level. They are the primary focus of data cleaning.

⁹² 80-20 Rule Definition (investopedia.com)

Table 6: Data quality problems classification in data sources

		Data quality problems	
		Single-source problems	Multi-source problems
Category	Schema level	<ul style="list-style-type: none"> • lack of integrity constraints; • poor schema design; • Uniqueness constraints; • Referential integrity 	<ul style="list-style-type: none"> • heterogeneous data models and schema design; • naming Conflicts • structural conflicts
	Instance level	<ul style="list-style-type: none"> • data entry errors; • misspelling; • redundancy/duplicates • contradictory values 	<ul style="list-style-type: none"> • Overlapping, contradicting and inconsistent data; • inconsistent aggregating; • inconsistent timing

3.3.4 Quality Dimensions of Linked Open Data

Data quality dimensions is a term used by data management professionals to describe a feature of data that can be assessed or measured by defined standards, allowing to determine the quality of data [211]. *Wang and Strong* (2013) used systematic approaches to identify and describe data quality. The authors identified three different approaches to study data quality: *the empirical, the theoretical, and the intuitive approach* to identify more than a hundred data quality dimensions important to data consumers.

The identified attributes are grouped into 20 data quality dimensions, each representing a single aspect of data quality[135][246]. The outcome of these approaches produced the definition of data quality dimensions from the following three perspectives [247]:

- 1. User perspective:** *defines data quality dimensions according to user's intended use and expectations;*
- 2. Data perspective:** *selects quality dimensions based on goals of the specific application and enabling an objective and automatic data quality assessment;*
- 3. Real-world perspective:** *presumes that an information system represents an application domain; derived from the theoretical approach that examines the origin of data deficiencies, and allows the definition of a comprehensive set of data quality dimensions [135].*

In data quality literature, several authors contributed to building an extensive list of data quality dimensions as in [248][135][249][250][251], (see Table 7 for the most cited dimensions, the categorization is according to Wang & Strong, 2012 [135]). A list of the most frequently discussed data quality dimensions with their criterion in literature is presented in Table 8.

Table 7: List of research articles discussed data quality dimensions

(Author, Year), Ref.	Category	Dimensions discussed
(Woodall et al., 2013), [252]	<i>Intrinsic</i>	Accuracy
	<i>Contextual</i>	Completeness
Woodall et al., 2014), [253]	<i>Contextual</i>	Completeness
Hazen et al., 2014), [254]	<i>Intrinsic</i>	Accuracy, Consistency
	<i>Contextual</i>	Timeliness, Completeness
(Kwon et al., 2014), [255]	<i>Intrinsic</i>	Consistency
	<i>Contextual</i>	Completeness
(Rao et al., 2015), [256]	<i>Intrinsic</i>	Accuracy, Timeliness
	<i>Contextual</i>	Confidentiality, Completeness, Volume
(Cai & Zhu, 2015), [132]	<i>Contextual</i>	Availability, Usability, Reliability, Relevance, and Presentation quality
(Serhani et al., 2016), [257]	<i>Intrinsic</i>	Accuracy, Consistency, Timeliness
	<i>Contextual</i>	Completeness
(Taleb et al., 2016), [222]	<i>Intrinsic</i>	Accuracy, Consistency
	<i>Contextual</i>	Completeness
(Taleb & Serhani, 2017), [258]	<i>Intrinsic</i>	Accuracy, Consistency
	<i>Contextual</i>	Completeness
(Xie et al., 2017), [259]	<i>Intrinsic</i>	Accuracy, Consistency
	<i>Contextual</i>	Completeness, Validity
(Zhang et al., 2017), [138]	<i>Contextual</i>	Availability, Usability, Reliability, Relevance
(Catarci et al., 2017), [260]	<i>Intrinsic</i>	Accuracy, Consistency
	<i>Representational</i>	Confidentiality
(Taleb et al., 2018), [261]	<i>Intrinsic</i>	Accuracy, believability, Consistency, Timeliness
	<i>Contextual</i>	Reputation, Relevancy, Value-added, Completeness
	<i>Representational</i>	Interpretability, Representational conciseness, Manipulability
	<i>Accessibility</i>	Access, Security, Ease of understanding
(Ardagna et al., 2018), [262]	<i>Intrinsic</i>	Accuracy, Consistency, Timeliness
	<i>Contextual</i>	Completeness, Volume, Distinctness, Precision
(El Alaoui, Gahi & Messoussi 2019), [263]	<i>Intrinsic</i>	Accuracy, Consistency
	<i>Contextual</i>	Completeness, uniqueness, freshness, transformation, conformity, normalization, referential integrity, credibility

Table 8: Most frequently used data quality dimensions along with their criterion

Category	Dimension	Brief Description	Criterion
<i>Intrinsic Category</i>	<i>Accuracy</i>	Conformity to the standards of admissible errors of numerical evaluation	<ul style="list-style-type: none"> • Syntactic validity of RDF docs. • Syntactic validity of literals • Syntactic validity of triples
	<i>Consistency</i>	Degree of certainty that data stored in distributed databases describing the same properties of the same objects and have the same values	<ul style="list-style-type: none"> • Check of schema restrictions during insertion of new statements • Consistency of statements w.r.t relations • constrains Consistency of statements w.r.t class constrains
	<i>Trustworthiness</i>	the degree of confidence in data, interpretation, and methods used to ensure the quality of a study	<ul style="list-style-type: none"> • Trustworthiness on KG level • Trustworthiness on statement level • Using unknown and empty values
<i>Contextual Category</i>	<i>Completeness</i>	The level on which the data contain all desired components	<ul style="list-style-type: none"> • Schema completeness • Column completeness • Population completeness
	<i>Timeliness</i>	The age of data	<ul style="list-style-type: none"> • Timeliness frequency of the KG • Specification of the validity period of the statement • Specification of the modification data of the statement
	<i>Relevance</i>	The level of data conformity to the user requirements	<ul style="list-style-type: none"> • Creating a ranking of statement
<i>Representational Category</i>	<i>Interoperability</i>	Data ability to be used and processed in a distributed group cooperation	<ul style="list-style-type: none"> • Avoiding blank nodes and RDF reification • Provisioning of several serialization formats • Using external vocabulary • Interoperability of proprietary vocabulary
	<i>Ease of understanding</i>	The level on which the data are understandable or interpretable by the user	<ul style="list-style-type: none"> • Description of resources • Labels in multiple languages • Understandable RDF serialization • Self-describing URIs

Accessibility category	<i>Accessibility</i>	The level on which the data are easily available or retrievable	<ul style="list-style-type: none"> • Dereferencing the possibility of resources • Availability of the KG • Provisioning of an RDF export • Provisioning of public SPARQL • Linking HTML sites to RDF serialization • Provisioning of KG metadata • Support of content negotiation
	<i>License</i>	The granting of permissions for a consumer to re-use a dataset under a defined condition	<ul style="list-style-type: none"> • Provisioning machine-readable licensing information
	<i>Interlinking</i>	The extent to which entities that represent the same concept are linked to each other be it within or between two or more data sources	<ul style="list-style-type: none"> • Interlinking via <i>owl:sameAs</i> • Validity of external URIs

A shorter list can be summarized based on frequently mentioned dimensioned in literature such as [211][248][223][264][265][266] [267], as follows:

1. **Accuracy** evaluates “the extent to which data is correct, reliable and certified free of error” [135], and could be calculated as the “quotient of the number of correct values in a source and the overall number of values” [250].
2. **Completeness** Takes into consideration if a dataset includes all data necessary to “represent every meaningful state of the represented real-world system” [248], and should consider why a value is missing [268].
3. **Consistency** Refers to “the violation of Semantic rules defined over a set of data items” [268] and “the extent to which data are always presented in the same format and are compatible with previous data” [135].
4. **Timeliness** influenced by system volatility (rate of change), currency (time of data update), and the time the data is used [248] and described e.g. as “the extent to which the age of the data is appropriate for the task at hand” [135] or “the average age of data in a source” [250].
5. **Relevancy** Evaluates whether available data types are pertinent to the intended use of the data [269] and described as “if the provided information satisfies the user’s need”[250].

It becomes apparent that key data quality dimensions are not universally agreed upon [270]; however, the International Data Management Association (DAMA) provides a modified comprehensive list of the data quality dimensions as represented in Figure 20 [Adapted from DAMA, 2013].



Figure 20: Data Quality Dimensions [Source DAMA, 2013]

DAMA (2013) defined the six data quality dimensions, in which they added *uniqueness* (nothing will be recorded more than once, based upon how that thing is identified), *validity* (data are valid if it conforms to the syntax (format, type, range) of its definition), and excluded *relevancy*[271]. Additional factors that can hinder the efficient use of data include usability, flexibility, confidentiality, and value timing issues [130],[131]. Conditions that contribute to data quality problems include *lack of validation routine* [135]; *correct but not valid data* [273]; *mismatched syntax, data formats, and structures* [274]; *unpredictable changes in the source system; a set of interfaces; absence of referential integrity checks; vulnerable system design, and; data conversion glitches*⁹³. Although the 5-star scheme of open and linked data presented by *Tim Berners-Lee* is widely cited shown in chapter 2, it only covers a specific data quality aspect, i.e., the format or encoding used to publish the data. This results that a dataset that can achieve the 5-star level while showing at the same time poor quality, such as data inaccuracy, inconsistency, and irrelevancy, etc. [134].

Understanding basic terms such as accuracy, consistency, and relevancy, used to describe quality measurements, has sometimes proven difficult, even within the analytical community, mainly because, same words being used with conflicting meaning, and the qualitative concepts of accuracy and consistency are well established in English and some other languages such as German, but relatively new in some other languages such as Arabic.

3.3.5 Challenges Facing Linked Data Quality Dimensions

In literature, numerous challenges facing linked data quality dimensions are summarized in [121][153][275][229]:

- *Linked data indicates a Web-scale knowledge base comprised of interlinked published data from several isolated information providers with different quality based on data provider objectives. The published data may contain incomplete or inaccurate metadata that affects the quality of the dataset.*
- *As data sizes increase, it becomes harder to assess its quality.*

⁹³ <https://slidewiki.org/deck/99262-1/big-data/slide/635258-2/635258-2:3/view>

- *The utilization of linked datasets by third-party applications may differ from dataset original creators' expectations.*
- *Linked data offers data integration via data interlinking among heterogeneous data sources. Integrated data quality is related to the original data sources quality, which cannot be directly modeled.*
- *Related linked data in some cases, may be regarded as a dynamic environment where data can alter rapidly and cannot be presumed to be static (velocity of data). Alterations in linked data sources should reflect changes in the real world; otherwise, data can soon become obsolete. Outdated data may express data inaccuracy problems and can deliver invalid data.*

3.3.6 Data Quality Dimensions Classifications Schemes

Data quality dimensions show a critical management element in the data quality domain. Researchers and practitioners have proposed numerous classifications of dimensions in data quality, many of which have overlapping, and sometimes conflicting interpretations. Despite the numerous classifications, few of which have concentrated to consolidate these viewpoints. The foundation for selection (or exclusion) of the classifications and their constituent dimensions has not been established. In literature, many classification schemes for data quality dimensions emerged over the years, among of which is:

1. *Wand and Wang (1996) categorize 26 quality dimensions using a theoretical approach [248],*
2. *Batini et al. (2009) compared several data quality classifications and concluded that "no general agreement exists either on which set of dimensions defines the quality of data, or on the exact meaning of each dimension" [268].*
3. *Zaveri et al. (2012) conducted a comprehensive survey on linked data quality assessment and identified 16 quality dimensions [276]. The dimensions are classified into four categories: accessibility, intrinsic, contextual, and representational. Furthermore, In this context, based on the user-driven quality evaluation survey of DBpedia conducted by Zaveri et al. (2013) as a centerpiece of the Linked Open Data Cloud, 17 data quality problem types and 58 users assessed a total of 521 resources identified [277].*
4. *Naumann (2002) selected 22 quality criteria in 4 categories with an evident approach supported by literature research [250] ;*
5. *Price and Shanks (2016) stated that quality criteria should not be based on a single approach but be "both theoretically and practical grounded", and proposed a framework with 16 data quality dimensions [269].*
6. *Hitzler et al. (2012) developed a comprehensive methodological quality assessment framework for linked data based on 18 quality dimensions and 69 metrics [267], [277].*
7. *Radulović et al. (2018) developed a linked data quality model based on this work and the ISO/IEC 25012:2008 data quality model [278], each of the many quality dimensions is linked to a specific metric.*

A comprehensive classification of the data quality dimensions is contributory in the chase of developing a rationalized and unified set of dimensions that can help in a shared comprehension within the wider community and offer a basis for modeling of data quality prerequisites.

3.4 Linked Data Quality Assessment Methodologies - Comparison

3.4.1 Definitions

A data quality assessment methodology can be defined as *the process of evaluating if a segment of data satisfies the information that consumers require in a selected use case* [141][281][308]. According to Bizer and Cyganiak (2009), data quality assessment methodology is the “*process of evaluating whether a piece of data meets the information consumers need in a specific use case*” [141]. Batini et al.(2009), presented an overview of existing methodologies available in [223]. The process involves measuring the user-relevant quality dimensions and comparing the results of the assessment with the user’s quality requirements. Data quality methodologies can be categorized based on some criteria such as:

- **data-driven vs. process-driven:**
 - *data-driven* strategy based on using data sources solely for data quality improvement. Related data-driven improvement techniques include acquisition of new data, standardization or normalization, error localization and correction, record linkage, data and schema integration, source trustworthiness, and cost optimization.
 - *Process-driven* a strategy where the data production process is analyzed and may be modified to identify and remove quality problems root causes. *The process-driven* strategy consists of two main techniques: process control and process redesign [223].
- **measurement vs. improvement:** used when measuring and assessing data quality is needed. Improvement and measuring procedures are closely interrelated. Assessment (benchmarking) is used when the measurements are compared to reference values to enable a diagnosis of dataset quality.
- **general-purpose vs. special-purpose:** *special-purpose* methodology focuses on a specific data domain, whereas the *general-purpose* methodology covers a wider spectrum of activities, domains, and phases.
- **Intra-organizational vs. inter-organizational:** benchmarking and improvement process covers a specific domain, sector, or organization. Alternatively, it concerns a group of organizations [251].

Diverse strategies are used by different methodologies to detect data quality problems, e.g., crowdsourcing mechanisms, manual, semi-automated, and automated approaches (as mentioned in section 3.5). Previous and existing researches and proposed practical solutions on data quality assessment can be categorized into the following groups:

1. *Definitions of data quality dimensions;*
2. *Guidelines of data quality;*

3. *Frameworks of data quality assessment;*
4. *Open data portals and/or open government data quality assessment; and*
5. *Linked data quality assessment*

3.4.2 Strategies and Techniques for Linked Data Quality Assessment

Linked Data quality assessment related works focused on the definition of metrics to quantify data quality according to different quality dimensions and designing a framework to provide tools supporting the calculation of the defined metrics. In literature, data quality is studied as a multi-dimensional concept [300][294][248][135]. Various techniques and approaches were developed to manage Linked Data quality aspects via introducing different systematic methodologies, the approaches can be generally classified into *manual, semi-automated, and automated*. The majority of early work on Linked Data quality was relevant to data trust. Among data trust-related works are:

- *Gamble and Goble (2011)* studied evaluating trust of Linked Data datasets [301];
- *Golbeck and Mannes (2006)* studies trust in networks based on the interchange of trust, provenance, and annotations [302];
- *Gil and Arts (2007)* studied the concept of reputation (trust) of Web resources [303];
- *Bonatti et al. (2011)* studied data trust based on annotations [41], *and;*
- *Shekarpour and Katebi (2010)* focus on the assessment of trust of a data source[304].

At a later stage, research work focused on other different issues of Linked Data quality such as accuracy, completeness, conciseness, consistency, dynamicity, relevancy, and accessibility, in particular, the work of:

- *Hogan et al. (2012)* presented a study focused on data quality assessment of main errors, noise, and modeling issues [305].
- *Lei et al. (2007)* studied some quality problems types related to accuracy. Especially, the evaluation of incompleteness, duplicate instances existence, ambiguities, and inaccuracy of instance labels and classification [284].
- *Rula et al. (2012)* focused on timeliness assessment aiming at reducing errors related to obsolete data. The authors defined the currency metric as the differences between the data current time and the last data modification time, also, they considered the time difference between observation and creation time of data [306].
- *Ellefi et al. (2018)* presented a comprehensive overview of the RDF dataset profiling feature, methods, tools, and vocabularies. The features of dataset profiling are organized into categories seven top-top-level: General; Qualitative; Provenance; Links; Licensing; Statistical; and Dynamics. For the qualitative features, the authors explored the data quality perspectives and outlined four categories: Trust; Accessibility; Representativity; and Context/Task Specificity [307].
- *Knuth et al. (2014)* identified the key challenges for Linked Data quality and outlined validation as one of the key factors in Linked Data quality. They stressed validation to be an integral part of the Linked Data lifecycle. The authors also outline the usage of popular vocabularies or manual creating of new correct vocabularies [275].

As Linked Data quality evaluation is gaining growing attention by the Semantic Web community. On the other hand, the current state of the art efforts is paying less attention to the understanding of knowledge base resource changes over time to detect abnormalities over various releases. As a result, defects still exist, such as: *unable to output easily explained results, user involvement in the process, applicable for selected Linked Datasets only, or evaluation of complete linked dataset during the assessment not possible.*

3.4.3 Comparison of Previous and Related Works

In literature, not many research works have dealt with linked data methodologies and linked open data life cycles, i.e., the process of generating, linking, publishing, and using linked data; to name a few: An introductory level guides *Bauer & Kaltenböck (2012)*, *Hyland & Villazón-Terrazas (2011)* [321]. Advanced “cookbooks” are the EUCLID curriculum⁹⁴, *Heath & Bizer (2011)* [175], *Morgan et al. (2014)*; *Ngonga Ngomo et al. (2014)* [283], *van Hooland & Verborgh (2014)* [322], *Atemezing et al. (2013)* [323], and *Wood et al. (2014)* [252]. The W3C Best Practices for Publishing Linked Data W3C-Government Linked Data Working Group (2014)⁹⁵ [324] include; A Cookbook for Publishing Linked Government Data on the Web [325]; Linked Data Life Cycles [326]; Guidelines for Publishing Government Linked Data [327]; Managing the Life-Cycle of Linked Data with the LOD2 Stack [328]; Methodological Guidelines for Consolidating Drug Data [143] are the highly cited best practices in the literature; see Table 11 below for a comparison [329].

Table 9: Comparison of linked data 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

⁹⁴ EUCLID - Educational Curriculum for the Usage of Linked Data, <http://euclid-project.eu>

⁹⁵ W3C Best Practices for Publishing Linked Data. <http://www.w3.org/TR/ld-bp/> (2018)

Guidelines for Publishing Government Linked Data		
Villazón-Terrazas et al. (2011)	(1) Specify, (2) Model,	Initialization
	(3) Generate, (4) Publish,	Innovation
	(5) Exploit	Validation & Maintenance
Managing the Life-Cycle of Linked Data with the LOD2 Stack		
Auer, et al. (2012)	(1) Extraction,	Initialization
	(2) Storage, (3) Authoring, (4) Interlinking, (5) Classification,	Innovation
	(6) Quality, (7) Evolution/Repair,	Validation & Maintenance
	(8) Search/ Browsing/ Exploration	Maintenance
Methodological guidelines for consolidating drug data		
Jovanovik and Trajanov (2017)	(1) Domain and Data Knowledge, (2) Data Modeling and Alignment,	Initialization
	(3) Transformation into 5-star Linked Data,	Innovation
	(4) Publishing the Linked Data Dataset on the Web,	Validation & Maintenance
	(5) Use-cases, Applications and Services	Maintenance

It is worth mentioning that the above life cycles (except the Linked Data Lifecycle proposed by *Auer et al. (2012)*) did not tackle the issue of data quality assurance or data repairing or cleaning at any of their work stages. This led to datasets that are concerned with generating data quantity at the expense of data quality as per studies [330][221][148][331][149][305]. *Auer et al. (2012)* proposed a lifecycle of Linked Data and dedicated two separate phases for quality assessment and repair out of the eight phases for the lifecycle [328].

One of the first linked data methodologies was developed in the European research project LOD2 (Creating Knowledge out of interlinked Data, 2011-2014)⁹⁶ that was mainly devoted to the publishing process, i.e., opening data in a machine-readable format and establishing the technologies and tools for integrating and interlinking heterogeneous data sources in general.

Jovanovik and Trajanov (2017) proposed methodological guidelines for consolidating drug data, they 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 modelling, the definition of the URI format for the entities and the ways of publishing the generated dataset....*” [143]. They also stated, “*The LOD2 tools are general, and cannot be applied in a specific domain without further work and domain knowledge....*”. Therefore, they proposed a new linked data methodology with a focus on reuse that provides guidelines for data publishers defining reusable components in the form of tools, schemas, and services for the given domain (i.e., drug management).

⁹⁶ <https://linkeddata.rs/project/LOD22010-2014>

Regarding useful tools such as converters for RDF, editors for Linked Data, RDF databases, etc. the W3C wiki offers an extensive tool directory (W3C wiki: Tools, <http://www.w3.org/2001/sw/wiki/Tools>). Some projects describe particular tools they endorse for different tasks of the Linked Data lifecycle, for example, the projects LATC (various tools)⁹⁷ and LOD2 (main tools of the project partners).

The methodology presented in this thesis is based on our new pilot application presented in [196], we developed an approach to convert and interlink drug data files created in some Arabic countries with selected datasets to enrich the knowledge and gain more value for end-users, concentrating on data quality for all phases of the process, the methodology comprises three phases and ten processes as presented in Table 12. The methodology will be discussed in detail in section 4.4.

Table 10: The proposed Linked Open Drug Data methodology

Methodological guidelines for quality assessment of Linked Data		
Guma Lakshen (2019)	(I) (1) Data Selection, (2) Data Analysis and (3) Data Cleaning, (quality assessment)	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 (for the overall process of the methodology), (V) (10) Use-cases, Applications and Services.	Validation & Maintenance

3.4.4 Comparison of Linked Data Quality Assessment Frameworks

In what follows, we present a comparison between some Linked data quality assessment frameworks and tools (see table 9 below), the comparison consists of Accessibility/availability, Extensibility, User interface, Automation, Licensing type, Collaboration, Customizability, Scalability, Usability, and Last Version release data. The compared tools are SWIQA, LODQM, LiQuate, TripleCheckMate, LINKQA, LUZZU, SIEVE, ODCleanStore, RDFUnit, ABSTAT, TrustBot, DaCura, ProLOD, tSPARQL, and TRELIS. Some of the techniques for quality assessment necessitate considerable manual efforts and don't scale up to huge dataset levels.

Research work on quality issues repairs is still inadequate compared to research studies on quality assessment. The following list presents several frameworks that already exist:

⁹⁷ LATC - LOD Around The Clock (EU, FP7-ICT, 9/2010-8/2012), <http://latc-project.eu>

Table 11: Comparison of existing data quality assessment frameworks and tools

Feature Tool	Accessibility/ availability	Extensibility	User Interface	Automation	Licensing	Collaboration	Customizability	Scalability	Usability	Last Version
<i>SWIQA</i> [281]	√	WIQA PL		Semi-automated	Apache V2	*	*	*	2	2009
<i>LODQM</i> [177]	√	Java /Jena API	√	automatic	-	*	*	*	3	2014
<i>LiQuate</i> [310]	√	Baysian rules	√	Semi-automated	-	*	*	*	1	2014
<i>TripleCheck Mate</i> [311]	√	*	√	Semi-automated	Apache	√	√	√	5	2013
<i>LINKQA</i> [312]	√	Java	*	automated	Open-source	*	*	*	2	2011
<i>Luzzu</i> [218]	√	Java, LQML	√	Semi-automated	Open-source	*	√	√	3	2016
<i>Sieve</i> [313]	√	XML	*	Semi-automated	Apache	*	*	√	4	2012
<i>ODCleanStore (ODCS)</i> [314]	√	Java	√	Automated, Semi-automated	Open-source Java	√	√	√	1	2012
<i>RDFUnit</i> [204]	√	SPARQL	*	Semi-automated	Apache	*	*	√	3	2016
<i>ABSTAT</i> [315]	√	SPARQL	√	automated	open-source GNU Affero GPL. v3.0	No	√	√	1	2015
<i>TrustBot</i> [316]	√	Java API	√	Semi-automated	-	No	√	No	4	2003
<i>DaCura</i> [317]	√	SPARQL	√	Semi-automated	Fuseki J. Apache	√	√	No	1	2013
<i>ProLOD</i> [318]	Screen-casts	-	√	Semi-automated	-	No	√	-	3	2010
<i>tSPARQL</i> [319]	√	SPARQL	√	Semi-automated	GPL v3	No	√	√	4	2012
<i>TRELLIS</i> [320]	-	*	√	Semi-automated	Open-source	√	√	*	2	2005

- *Fürber and Hepp* (2011) designed **SWIQA**, a framework applicable for Semantic Web resources as well as for relational databases with the support of wrapping technologies, such as D2RQ⁹⁸ platform. SWIQA identifies and classifies data quality problems, and calculates task-dependent and task-independent information quality scores using a quality rule template[281].

⁹⁸ It is a system for accessing relational databases as virtual and read-only RDF graphs, It offers RDF-based access to the content of relational databases without having to replicate it into an RDF store.

- *Behkamal et al.* (2014) proposed **LODQM**, a metric-driven framework that assesses dataset quality before publication on the LOD cloud, suitable for assessing schema and property completeness. It follows the Goal-Question-Metric approach⁹⁹ used to solicit metrics to assess datasets [309][177].
- *Ruckhaus et al.* (2014) illustrated **LiQuate**, which combines Bayesian Networks and rule-based systems to analyze data quality and links in the LOD cloud. It identifies ambiguities among the linked data and suggests possible inconsistencies and incompleteness. It was built on top of the Biomedical linked datasets that maintain data related to clinical trials, interventions, drugs, conditions, diseases, and their relationships. LiQuate utilizes visualization services implemented by the D3.js JavaScript library¹⁰⁰[310].
- *Kontokostas et al.* (2013) developed **TripleCheckMate**, a crowdsourcing technique tool used to estimate linked open data quality. The tool can be configured to assess any Linked data dataset using different taxonomies of quality issues. It permits human contributors to select RDF resources, recognizes issues related to RDF triples of the resources, and classifies them according to a pre-defined taxonomy of data quality problems. It was developed under the DBpedia data quality project as an application for evaluating DBpedia correctness. It only records the triples that are identified as 'incorrect' [311].
- *Gueret et al.* (2012) presented **LINKQA**, an extensible framework that measures the assessment of Linked Data mappings using network metrics. LinkQA provides real-time statistics of link quality parameters along with graphical visualization. The framework consists of five components (Select, Construct, Extend, Analyze, and Compare) assembled in a workflow form[312].
- *Debattista et al.* (2016) proposed **Luzzu**, a conceptual methodology for assessing Linked Datasets based on the data quality lifecycle. The original Luzzu UI only displays the quality score and does not display the identified problems to the user nor assist the user to understand the identified problems. Even with some limitations, the Luzzu framework allows users to easily analyze data quality from a single visual entry point. [218].
- *Mendes et al.* (2012) developed **Sieve**, a framework for expressing quality assessment and fusion methods flexibly, integrated into the Linked Data Integration Framework (LDIF), which handles data access, Schema mapping, and identity resolution, all vital preliminaries for quality assessment and fusion. The score calculation is based on concepts such as assessment metric, aggregate metric, data quality indicator, or scoring function[313].
- *Knap et al.* (2012) presented **ODCleanStore**, a framework enabling management of Linked Data: data cleaning, linking, transformation, and quality assessment. The tool provides data consumers with the possibility to consume integrated data, which reduces the costs of Web application development. It allows the aggregation of linked open data

⁹⁹ Is based upon the assumption that for an organization to measure in a purposeful way it must first specify the goals for itself and its projects, then it must trace those goals to the data that are intended to define those goals operationally, and finally provide a framework for interpreting the data with respect to the stated goals.

¹⁰⁰ D3.js - Data-Driven Documents

with the evaluation of the quality. The evaluation is carried out at the query time, integrating a phase of the resolution of conflict in the case of the presence of contradictory data [314].

- *Kontokostas et al.* (2014) developed **RDFUnit**, a pattern-based evaluation scheme for Linked Data quality assessment. It is a test-driven data-debugging framework able to run automatically generated (based on a schema) and manually generated test cases against an endpoint. It uses data schema and quality patterns created from DBpedia user community feedback, Wikipedia maintenance system, and ontology analysis. It assists in defining quality test patterns using SPARQL query templates. All test cases are executed as SPARQL queries using a pattern-based transformation approach [204].
- *Spahiu* (2015) developed **ABSTAT**, an ontology-driven linked data summarization model proposed to mitigate the data set understanding problem. ABSTAT framework enables users to query (via SPARQL), to navigate the summaries through Web interfaces. ABSTAT allows the use of data profiling and data mining techniques to explore linked data and to detect quality issues at the schema level [315].
- *Golbeck et. al* (2003) developed **TrustBot**, an IRC bot¹⁰¹ that makes trust recommendations to users (based on the trust network it builds), the users have the flexibility to submit their URIs to the bot at any time while incorporating the data into a graph. The bot retains a collection of these URIs that are spidered when the bot is launched or called upon to reload the graph. From an IRC channel, the bot can be queried to provide the weighted average, as well as max and min path lengths, and max and min capacity paths. The TrustBot is running on <http://trust.mindswap.org/trustMail.shtml> and can be queried under "TrustBot" [316].
- *Feeney et. al* (2014) developed **DaCura**¹⁰² framework aimed at providing dataset curators with tools to collect and curate evolving linked data datasets that maintain quality over time. The framework is designed to support harvesting, assessment, management, and publication of high-quality Linked Open Data. It requires a lot of human efforts for modifying schema involving domain experts, data harvesters, and consumers [317].
- *Böhm et al.* (2010) developed **ProLOD** is a Web-based tool that analyzes the object values of RDF triples and generates statistics upon them such as datatype and patterns distribution. In ProLOD the detection type is performed using regular expression rules and normalized patterns are used to visualize huge numbers of different patterns. ProLOD also generates statistics on literal values and external links. ProLOD++¹⁰³ which is an extension of ProLOD is also a browser-based tool that implements several algorithms intending to compute different profiling, mining, or cleansing tasks. In the profiling task, processes are included to find distribution and frequencies of subjects, predicates, and objects, range of the predicates, etc. ProLOD++ can also identify predicates combinations that contain unique values as key candidates to distinctly

¹⁰¹ An IRC bot is a set of scripts or an independent program that connects to Internet Relay Chat as a client, and so appears to other IRC users as another user

¹⁰² Documentation, demonstrations, and examples for the DaCura system is available at <http://dacura.cs.tcd.ie>

¹⁰³ <https://www.hpi.uni-potsdam.de/naumann/sites/prolod++/#/graphstatistics/dailymed>

- identify entities. The tool performs some cleansing tasks such as auto completions of new facts for a given dataset, ontology alignment in identifying predicates that are synonym, or identifying cases where the pattern usage is over specified or underspecified [318].
- *Hartig* (2008) proposed **tSPARQL** a trust-aware query language that enables SPARQL to designate trust requirements and access the query solutions' trustworthiness through redefining SPARQL algebra such that the resulting algebra operates over sets of trust-weighted valuations, that is, conventional SPARQL valuations that are associated with a trustworthiness score. tSPARQL adds two new operators that enable users to describe trustworthiness requirements and to access the trustworthiness of (intermediate) solutions; the latter may be used to obtain a trustworthiness-related ordering of a query result or to output trustworthiness scores as part of a query result. Similar to SPARQL, tSPARQL is defined for expressing queries over fixed, a-priori defined collections of RDF data (for tSPARQL these collections need to be augmented with a trust function) [319].
 - *Gil and V. Ratnakar* (2002) developed **TRELLIS** an interactive tool that aids users annotate the rationale for their decisions, hypotheses, and opinions as they analyze information from various sources. TRELLIS generates annotations of the user's analysis in several markup languages (XML, RDF, and DAML+OIL). TRELLIS has more support for assessing sources, sharing, and collaboration, but does not provide as much support for automation nor domain-specific standard patterns to facilitate sharing [320].

There is substantial work in the Semantic Web community to assess the quality of Linked Data. However, in the current state of the art, less focus has been given toward understanding knowledge base resource changes over time to detect irregularities over various releases, which is instead the main contribution of our approach.

3.5 A Conceptual Methodology for Linked Data Ecosystems Quality Assessment

According to *Bizer et. al* (2009), a data quality assessment methodology is defined as the process of evaluating if a piece of data matches the information consumers require for a specific use case [141]. *Zaveri et. al* (2012), observed that in all the 30 identified approaches, no standardized phases were followed for dataset quality assessment [276].

3.5.1 Proposed Methodology

Taking into consideration past methodologies, we suggest a methodology consisting of three phases and nine steps for assessing linked open data quality aims at supporting the assessment and evaluation of the quality of linked open data sources throughout the various stages of the data integration process and it consists of three main phases and six steps (see Figure 21) described as follows:

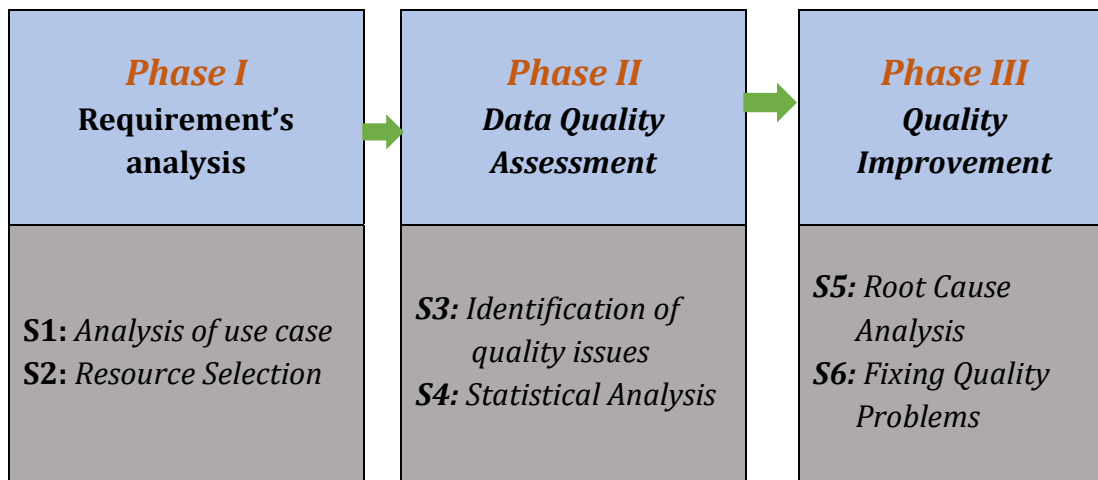


Figure 21: Methodology for Assessing Linked Open Data Quality

Phase I: Requirement's analysis

Step 1: Analysis of use case

- Collecting of requirements and subsequent analysis of the requirements based on the use case.
- Identifying user dataset requirements related to the use case in mind.

Step 2: Resource selection

- Selecting datasets required for quality assessment.

Phase II: Data Quality Assessment

- Select the most relevant dimensions and metrics are.
- Perform a quantitative evaluation of the quality of the dataset using the selected metrics specific for the selected dimension. Therefore, the phase includes:

Step 3: Identification of quality issues

- identifying a set of the most relevant data quality issues based on the use case.

Step 4: Statistical Analysis

- Performs basic statistical and low-level analysis on the dataset (e.g., number of blank nodes, number of interlinks between datasets).
- Calculate generic statistics on the dataset based on certain pre-defined heuristics.
- Overall assessment of the overall quality of the dataset through evaluating the results performed between target and original datasets or those in the same domain, and aggregating value (score) of the results.

Phase III: Quality Improvement

This phase emphasises improving the quality of the dataset based on the analysis achieved in Phase II focusing on the use case identified in Phase I.

Step 5: Root Cause Analysis

- discover the cause of the detected data quality issues i.e. execute root cause analysis to detect whether the problem occurs in the original dataset

Step 6: Fixing Quality Problems

- Identify strategies to address the identified root cause of the problems are implemented, such as Semi-automatic, automated approaches, or crowdsourcing approaches.

The above phases and their subsequent steps can be translated into a flow chart (see Figure 22) that shows an efficient data quality assessment process with a dynamic feedback mechanism based on big data’s characteristics.

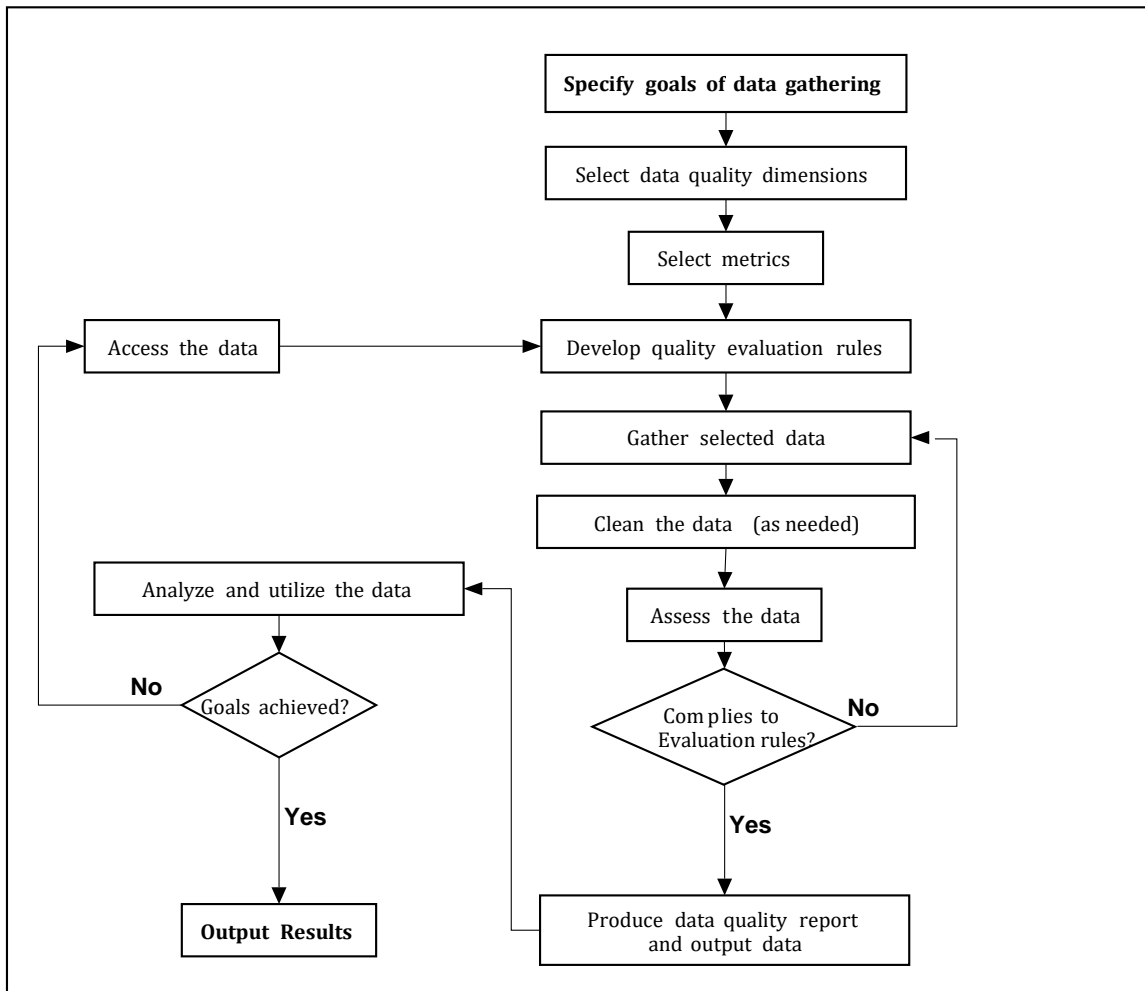


Figure 22: A Flow Chart of the Methodology for Assessing Linked Open Data Quality

A Data Quality Assessment is a distinct phase within the data quality life-cycle that is used to verify the source, quantity, and effect of any data items that violate pre-defined data quality rules. The first step of the process is to specify the data gathering goals of the assessment process. Users of big data usually select their data based on their requirements, such as operations, decision making, and planning. Then, selecting data quality dimensions is performed, where each quality dimension needs different measurement tools, techniques, and processes, which leads to differences in assessment times, costs, and human resources followed by metrics selection. After the completion of quality assessment preparation, the process enters the quality evaluation rules development phase. Big data sources are very wide and data structures are complex. The received data may have quality problems, such as missing

information, data errors, noise, inconsistencies, etc. Data cleaning (data scrubbing) purpose is to spot and remove inconsistencies and errors from data to improve its quality. Data cleaning may be divided into four patterns according to implementation methods and scopes as manual implementation, writing of special application programs, data cleaning unrelated to specific application fields, and solving the problem of a type of specific application domain. In these four approaches, the third has good practical value and can be applied successfully.

Then, the process enters the data quality assessment and monitoring phases. The core of data quality assessment is how to evaluate each dimension. The current method has two categories: qualitative and quantitative methods. Therefore, objectivity, generalizability, and numbers are features often associated with this method, whose evaluation results are more intuitive and concrete. After the assessment, the data can be compared with the evaluation rules. If the data quality complies with the evaluation rules standard and a data quality report will be generated. Otherwise, if the data quality fails then it goes back to gather new data.

If the analysis results encounter the goal, then the results are outputted and fed back to the quality assessment system to provide improved support for the next round of assessment. If results unmatched the goal, the data quality assessment baseline may not be reasonable, and we need to adjust it in a timely fashion to get results in line with our goals.

The process of creating high-quality statistics and reports relies highly on data quality assessment at all stages of data projects. Without a regular assessment of data quality, the outcomes of the reporting system will jeopardize the many statistical processes such as data gathering, editing, or weighting. Neglecting data quality assessment would wrongly assume that the processes are unimprovable and that errors will continuously be detected deprived of systematic analysis. Simultaneously, data quality assessment is a prerequisite for notifying the users about the probable utilization of the data, or which results are publishable. Certainly, ignoring respectable approaches for data quality assessment leaves statistical institutes working blindly and unable to claim that their work meets the quality requirement and professional norms. Data quality assessment is a process for evaluating if data fulfills the user's defined needs [5], whereas, data quality assessment methodology is frequently defined as *the process of evaluating if a portion of data meets the information consumer's requirements in a specific use case* [141]. The assessment is usually implemented using a data quality assessment framework.

In literature, several methodologies, tools, and metrics to evaluate data quality, in general, were developed. An inclusive survey conducted by *Zaveri et. al* (2016), stated that in the 30 identified approaches, there was no standardized set of phases that were followed to assess the quality of a dataset [5]. While such approaches assist as the guiding background knowledge for data quality measurement on the LOD, their implementation is not direct because data quality on the LOD is related to novel aspects such as data representation quality or consistency concerning the information existing in other published datasets. Furthermore, mechanisms of knowledge inference on the LOD frequently follow an open world assumption, while the existing methods generally assume closed world Semantics.

The dimensions are indirectly scaled using one or more quality metrics. These metrics submit several values (typically normalized between 0 and 1) which can then be compared to desired thresholds for accept/reject quality assessment or observing quality attitudes over time (see section 3.3.3 above). Some assessment frameworks might also generate problem reports for the corrupted or missing data detected during calculating the metrics procedure. A single defect in a dataset could create a consecutive report as various metrics are assessed. To improve data quality, data corrections procedures should be implemented.

The user implementing the data correction must have a clear understanding of which quality problems (flaws) are present in the data, how to fix them, where they occur in the dataset, which metrics are impacted by each defect, and how much improvement each fix would bring. Data quality is usually an expensive process and thus a preference for defect fixing is important as there is a trade-off between cost and quality.

3.5.2 Selection of Data Quality Dimensions for quality evaluation rules

Data quality assessment is the process of testing the data against a subset of quality indicators, resulting in a fixed value used to check if the data meets the required quality. To accomplish the measurement of quality data, it is important to make assessments of some related dimensions [213]. Generally, most data quality assessment attributes depend on user experience which could be dependent on user perception, and other attributes are linked with the data itself. As for this thesis concerns, we selected three dimensions accuracy, consistency, and relevancy from the most used dimensions selected (see Table 13 above), to be studied and analyzed.

a. Accuracy

In literature [135], the accuracy dimension determines the extent to which data are correct, reliable, and certified free of error. Accuracy indicates the extent to which entities and facts truly represent the real-life phenomenon. In this sense, accuracy is assessed by comparing data with their sources in reality. For example, data accuracy refers to *“the degree with which data values agree with an identified source of correct information”* [224][279]. Thus, accuracy in this sense is pertinent to the process of data creation. Likewise, to relational data, accuracy in linked data could be classified into syntactic and semantic accuracy.

- **Syntactic Accuracy:** *Peralta (2006)* defined it as: *“data is argued to be correct, in a syntactic way, if it satisfies syntactic rules and constraints imposed by the users”*. It is also defined as, *“the closeness of the data values to a set of values defined in a domain considered syntactically correct”*[280]. *Rula et al. (2016)* defined it as *“the degree to which an entity document conforms to the specification of the serialization format and literals are accurate for a set of syntactical rules”*[127]. Syntactic accuracy problems usually refer to *literal’s incompatible with data type range* or *malformed data type literals*. Misspelled literals can be considered as syntactic inaccurate data [149], for example, the words *theater*, *catalogue*, and *fulfil* are misspelled literals concerning *theatre*, *catalog*, and *fulfill* respectively. Regarding software tools, validators are used to detect syntactic accuracy concerning data types (measured in terms

of correct/incorrect values for a given property), ranges (measured in terms of correct/incorrect value range for properties holding numerical values), and syntactic rules (correct/incorrect values concerning given patterns) [149] [281].

• **Semantic Accuracy:** It refers to the accuracy of the meaning. Salgé (1995), defined it as *“The purpose of Semantic Accuracy is to describe the Semantic distance between geographical objects and the perceived reality”* [282]. Alternatively, the W3C, defined as *“the degree to which data values correctly represent the real-world facts”*. Semantic accuracy is more difficult to assess than syntactic accuracy because the vocabulary containing the definition of all terms in the syntactic accuracy is sufficient for the metric assessment. In literature, some metrics are proposed as follows:

- **validity of a fact** that checks the Semantic accuracy of the fact against several sources or even several Websites [283];
- **accuracy of the annotation**, representation, labelling, or classification that is detected as a value between 0 and 1 [284];
- **The semantic accuracy of the dataset** can be verified with the help of an unbiased trusted third party (humans) [228].

Generally, accuracy refers to the degree to which the data is correct, reliable, certified, and free of error [135]. The data requires to reflect the actual state of user expectations in terms of real-world representation through data acquisition and processing.

b. Consistency

Consistency refers to *“the degree to which the data is presented in a format that is the same and compatible with previous data”* [135]. Consistency implicitly implies that *“two or more values do not contradict with each other”* [228]. In general terms, it is defined as being free of conflicting information, although consistency doesn't necessarily mean correctness. Loshin (2006) defined it as, *“... in its most basic form, consistency refers to data values in one dataset being consistent with values in another dataset”* [285]., *System Analysis Program Development SAP*, also defined consistency in a similar approach [286]. The Health Information and Quality Authority¹⁰⁴ HIQA in [287], defined it as *“Comparability of data refers to the extent to which data is consistent between organizations and over time allowing comparisons to be made”*. This definition emphasizes that data should be consistent between the organizations to make comparisons. However, consistency can refer to several data aspects. For example, for data value: the value or entries in the data should be the same in all cases; concerning data representation: the entity types and attributes should have a similar basic structure wherever possible. The consistency of record fields depends on whether they follow a consistent syntactical format, without contradiction or discrepancy within the entire catalogue of metadata [288][289]. Regardless of the syntactical format, a field is regraded to be consistent if the relevant values are selected from a fixed set of options. An example of inconsistency is if within two records, the use of “U.A.E” and “United Arab Emirates” is

¹⁰⁴ Health Information and Quality Authority (HIQA) (healthcomplaints.ie)

interchangeable. Another example is the date representation order, e.g., day/month/year, or month/day/year, or any other arbitrary order.

c. Relevancy

Relevancy is the extent to which information is applicable and helpful for the task at hand [265]. Relevancy is regarded as a significant quality dimension in the Web-based system's domain, as information consumers are often faced with an excess of potentially relevant information. Relevancy may refer to the provision of information which is per the task at hand and important to the users' query. In literature, there exist many diverse definitions of relevancy, such as:

- *Wang and strong (2013) defined relevancy as "Data are applicable and useful for the task at hand" [135].*
- *Sowey and Petocz (2016) defined it as "Relevance is the degree to which statistics meet current and potential users' needs. It indicates that whether all the needed statistics are produced and the extent to which concepts used (definitions, classifications, etc.) indeed reflect user needs" [290].*
- *Health Information and Quality Authority HIQA defined it as "Relevance of data refers to the extent to which the data meets the needs of users. Information needs may change and are important that reviews take place to ensure data collected is still relevant for decision-makers." [287].*
- *Stvilia et al. (2007) defined it as "The extent to which information is applicable in a given activity" [291].*

According to the above definitions and many others, we may define the relevancy dimension as *"The Characteristics in which the Information is the valid type of information which adds value to the current task, to perform a process or aid decision-making"*. Relevancy is extremely context-dependent and is highly recommended in Web information systems as the process of retrieving the related information becomes sophisticated when dealing with big information flow. As an example of relevancy, let's consider a person is looking for information about a particular medicine and seeking 'relevant information, i.e., side effects/contradictions/interactions/ administration during pregnancy, etc.,

Most of the available commercial websites and datasets embed 'irrelevant information' as doctors, hospitals, clinical information, etc., in addition to the desired relevant information, and as a result, much irrelevant extra information is made available to the user, which may divert the passenger main attention. Providing *irrelevant data* deflect application developers and potential users and wastes network resources. Instead, restricting the dataset to only *relevant* information simplifies application development and increases the likelihood to return only relevant results to users. The retrieval process of relevant data can be performed through:

- *using a combination of hyperlink analysis and information retrieval methods [228];*
- *ranking (a numerical value similar to PageRank, which determines the centrality of RDF documents and facts [41]);*

- *counting the occurrence of relevant data within metadata attributes (e.g., title, description, subject) [228].*

Approaches to assessing Web document's relevancy are used within Web search engines, which sort documents based on their relevancy for a given query using a combination of hyperlink analysis [292] and information retrieval methods [293]. An alternative metric could be the coverage (i.e., number of entities described in a dataset) and level of detail (i.e., number of properties) in a dataset to ensure that there exists an adequate suitable volume of relevant data for a specific task [131].

3.6 Summary

The main function of a data ecosystem is to capture data and produce useful insights and value. Hence, nowadays, organizations are increasingly depending on data analysis to gain data value and achieve a competitive advantage. As data size gets bigger, creating a real value from such big data is possible if data passes quality assessment tests. Fulfilment of dimensions such as *accuracy, completeness, consistency, relevancy, and reliability of data* is essential to make good decisions and actions [131]. To guarantee that data conforms with an acceptable level of quality, methods and techniques performing data quality assessment are obligatory to support the identification of suitable data to process [132].

In this chapter, we proposed a **Conceptual Methodology for Linked Data Ecosystems Quality**. Taking into consideration past methodologies, we suggest a methodology consisting of three phases and nine steps for assessing linked open data quality that aims at supporting the assessment and evaluation of the quality of linked open data sources throughout the various stages of the data integration process. Generally, the design of a data quality assessment process depends on user experience which could be dependent on user perception and other attributes linked with the data itself. As for this thesis concerns, we selected three dimensions accuracy, consistency, and relevancy from the most used dimensioned selected (see Table 13) above, to be studied and analyzed with datasets from Arabic countries.

The comparison of related methodologies and tools was presented at ICIST 2019

- *Guma Lakshen, Valentina Janev, and Sanja Vraneš. 2019. Quality Issues of Open Big Data Ecosystems: Toward Solution Development. In: Konjović, Z., Zdravković, M., Trajanović, M. (Eds.) ICIST 2019 Proceedings.*

The Conceptual Methodology was presented at CISIM 2019

- *Guma Lakshen, Valentina Janev, and Sanja Vraneš. 2019. "Linking Open Drug Data: Lessons Learned". IFIP International Conference on Computer Information Systems and Industrial Management. book: Computer Information Systems and Industrial Management. (pp.164-175). DOI:10.1007/978-3-030-28957-7_15.*

Chapter Four

“Toward Solution Development” for Consolidating Arabic Linked Drug Datasets

CHAPTER FOUR – “TOWARDS SOLUTION DEVELOPMENT” FOR CONSOLIDATION OF ARABIC LINKED DRUG DATASETS

4.1 Introduction

In this chapter we will address and explain the process of creating the consolidated linked Arab drugs dataset and development of the Arabic Linked Drug Data Applications (ALDDA) through the following sections:

Section 4.2 discusses the main concepts related to the development of the ALDDA Knowledge graph using the RDF data model and points out the relationship between knowledge graphs and ontologies.

Section 4.3 introduces the most notable and existing Arabic drugs-related datasets on the Web. The linked open drug data LODD is also discussed in this section.

Section 4.4 discusses the requirements, datasets, and tools for the interlinking and enhancements of the Arabic datasets.

Section 4.5 introduces in detail the general phases of the ALDDA piloting methodology.

Section 4.6 studies the Selection and Implementation of Data Quality Assessment Measures and the Data quality Assessment functional forms.

Section 4.7 validates the ALDDA approach by discussing the implementation phases and processes of ALDDA which includes data selection, data analysis, data cleaning, mapping schema, ontology definition, data interlinking through reconciling with DBpedia and DrugBank, data publishing, and storing.

4.2 Towards the development of a Knowledge Graph

4.2.1 Using the RDF data model

RDF provides a standardized manner in expressing information such that it can be passed over between various systems preserving the same meaning [47]. The resource is a *thing* that can be referenced by a URI, the RDF is an adequate means to describe that *thing* any type even when it is inaccessible directly from the Web [47]. Some of the basic concepts and terminology used by RDF are dataset, document, triple, term, and serializations, more details can be found in [48].

To understand the mechanism of how the RDF data model works. Consider the following two statements:

- **“Pfizer-BioNTech treats Covid-19 infection”**
- **“Covid-19 infection treated by Pfizer-BioNTech”**

The two statements have the same meaning (i.e., **Pfizer-BioNTech** *treats* **Covid-19 infection**) for us as humans, but for computers, the above statements have different string structures which means that they are not the same. To remove this obvious ambiguity, the RDF data model needs the declaration of a resource *Web*. Thus, the data model corresponding to the statement " **Pfizer-BioNTech** *treats* **Covid-19 infection**" has a single resource **Pfizer-BioNTech**, a property-type of the *treats*, and a corresponding value of **Covid-19 infection**, (see Figure 23).

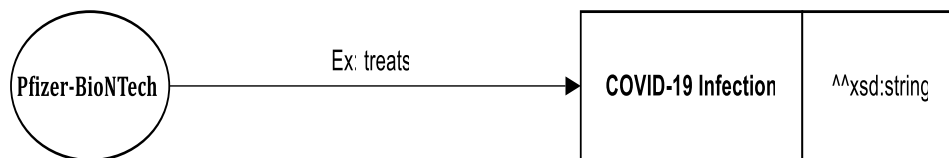


Figure 23: RDF graph example

The above RDF graph can be written as an RDF triple:

Ex: Pfizer-BioNTech Ex:treats "COVID-19 Infection" ^^xsd:string

Additional information for Covid-19 infection such as where it is identified, place, data causes, means of transmission, and symptoms, etc., to the example above RDF graph as required (see Figure 24).

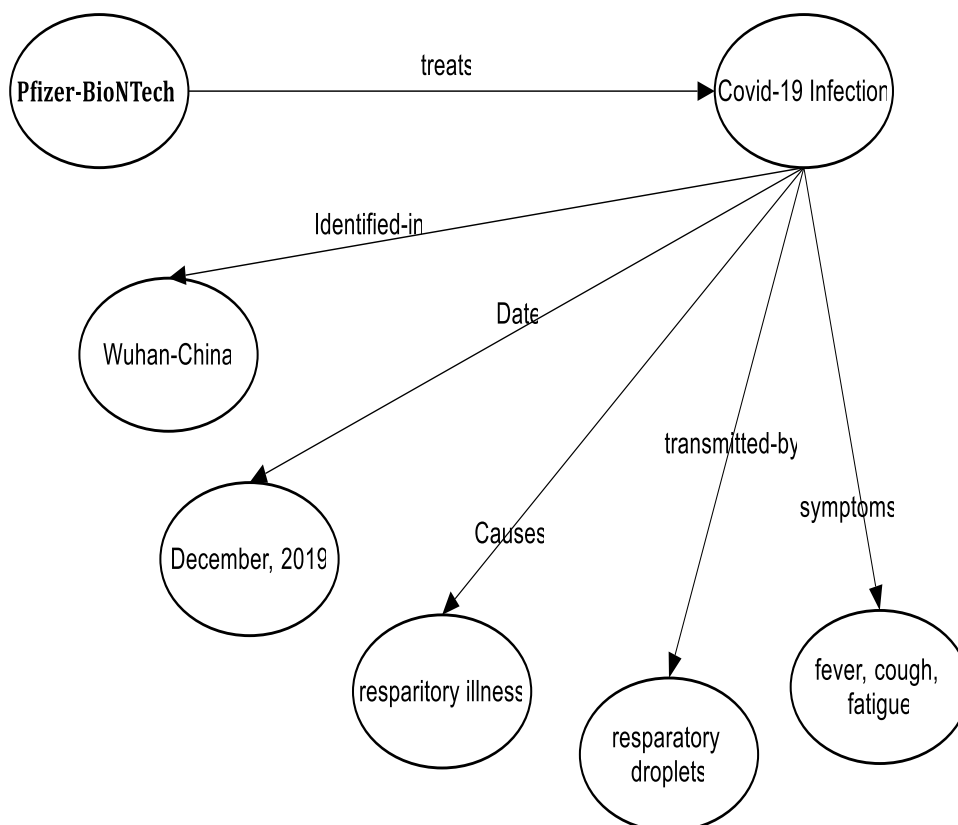


Figure 24: RDF extended example

The above RDF graph described can be written in the Turtle syntax as follows:

@ prefix	ex : <http:// example.org/ontology/> .	
@ prefix	ex : <http:// www.w3.org/1999/02/22-rdf-syntax-ns#>.	
@ prefix	ex : <http:// www.w3.org/2000/01/rdf-schema/#>.	
ex : Covid-19	rdf : type	ex : infection
ex : Covid-19	ex : indentified-in	ex : Wuhan-China
ex : Covid-19	ex : Indentified-on	ex : December, 2019
ex : Covid-19	ex : causes	ex : resparatory illness
ex : Covid-19	ex : transmitted-by	ex : resparatory droplets
ex : Covid-19	ex : symptoms	ex : fever, cough, fatigue

In RDF, Datatypes follows existing XML Schema standard which defines a hierarchy of data types along with their syntax [50], and the language tagged strings in RDF should be defined following RFC 3066 [51], to express the phrase “Semantic Web” in multiple languages:

Ex: SWeb

```

rdfs: label "Semantic Web" @en;
rdfs: label "semantički Web" @sr;
rdfs: label "الويب الدلالي" @ar;

```

To describe resources RDF establishes a set of terms, the most relevant is the **rdf:type** which is used to declare that a resource is a member of a defined class.

4.2.2 Ontologies and Knowledge Graphs

- **Ontology**, Linguistically, is a combination of two Greek words (*onto* means being and *logos* means study) and is rooted in a philosophy where it refers to the subject of being, existence, and basic categories [55]. An ontology is a set of concept definitions and relations between the concepts. It can be used to define what entities exist and also what entities may exist within a domain, see for instance Figure 14 illustrates the relationship between the concepts (nodes). Generally, Ontologies is defined as a "*representation of a shared conceptualization of a specific domain*"¹⁰⁵. Ontologies are the foundations of the Semantic Web and linked data as it specifies the shared knowledge and exchanges it between different systems. The knowledge specified can be defined through the Semantics of the utilized terms for describing data and the relations between these terms. Ontology in Semantic Web and computer science is recognized as a formal representation of knowledge [57]. For metadata development, an ontology is only obtained after defining all data elements, controlling and fitting vocabularies together.

¹⁰⁵ What is Ontology | IGI Global (igi-global.com)

Ontologies are important for the Semantic Web, although they don't have a specific definition but can be considered as a group of URIs. On the Semantic Web, it is the concepts and relationships which describe a range of concerns, also it is used to describe the complex [58].

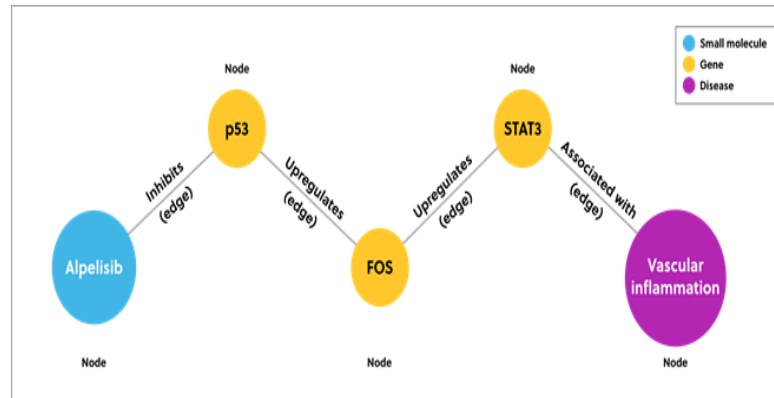


Figure 25: Example of connections between data using nodes and edges. Source [Exploring Knowledge Graphs for COVID-19 Drug Discovery | CAS]

OWL is reportedly most popular among the W3C standard language because of its expressiveness. There are several ontology editor tools such as Protégé¹⁰⁶, FAO AGROVOC Concept Server Workbench Tool [59]; OBO-Edit [60]; SWOOP [61]; Apollo [62]; IsaViz¹⁰⁷; TopBraid Composer [63], Jena [64], SESAME [65], KOAN[66], etc. A comparison of Tools, Languages, and Formalisms for ontology development can be found in [67][63].

- **Knowledge Graphs** The term knowledge graphs (KGs) is often used to refer to knowledge bases in the semantic web context [68]. Google introduced Knowledge graphs (KG) in 2012, as an unformal representation of interlinked data which substantially improves the search queries¹⁰⁸. In KGs, labelled concepts are represented by nodes and the edges represent semantic relations between nodes. KGs are defined as “*data storage structures that depend on principles from graph theory to represent information*”. Facts are stored as triples that bring together two entities through a relation. In a graphical context, these entities are identical to nodes, and the relations between them are identical to edges [69]. Ontologies are frequently used in association with knowledge graphs to offer an axiomatic foundation on which knowledge graphs are constructed. Public knowledge bases e.g., DBpedia¹⁰⁹, YAGO¹¹⁰, and WikiData¹¹¹ are all anchored by large-scale knowledge graphs including over one billion triples each. Google, as an

¹⁰⁶ [protégé \(stanford.edu\)](http://protégé.stanford.edu)

¹⁰⁷ [IsaViz Overview \(w3.org\)](http://IsaViz.org)

¹⁰⁸ [Introducing the Knowledge Graph: things, not strings \(blog.google\)](http://Introducing the Knowledge Graph: things, not strings (blog.google))

¹⁰⁹ www.DBpedia.org

¹¹⁰ [Home | Yago Project \(yago-knowledge.org\)](http://Home | Yago Project (yago-knowledge.org))

¹¹¹ Wikidata

example, uses a KG derived from Freebase¹¹² to reinforce their computer program results by providing info-boxes that summarize facts a couple of user's queries [70].

A knowledge graph is created when an ontology (data model) is populated with the targeted and harmonized heterogeneous data coming from different data sources. Manual graph construction is time-consuming and demands curators knowledgeable within the field. DBpedia, for example, relies on its community to curate its class taxonomy. In the next subsections, we will show how the Arabic datasets were integrated and interlinked with other public data.

4.3 Selection of LOD and Arabic Linked Drug Datasets

The linked Data concept allows users to traverse big volumes of varied data that existed on distributed sites on the Web, by starting at one single point. This permits the formation of use-case scenarios that deliver the end-users with added information and services, previously unobtainable over the isolated datasets. In this part of the thesis, we will deliver a process for Arabic Linked Drug Data Application, we named ALDDA as a use-case scenario to illustrate the competencies of the Linked Data nature and its benefits in a multilingual environment.

4.3.1 Existing Arabic Drugs-related Datasets on the Web

Utilizing the internet to dig for drug-related information has gained more attention over the years and has become a common practice worldwide. In the Arabic speaking region, there are only a few Arabic drug applications such as Webteb¹¹³, Altibbi¹¹⁴, 123esaaf¹¹⁵, and Kuwait Pharmacy KP¹¹⁶, Epharmapedia¹¹⁷, Dawee¹¹⁸, أدوية.كوم - adwyaa.com¹¹⁹, etc., which provide their services in Arabic and English languages, but unfortunately, their data are not open, not updated regularly, and mostly not free. Some of the most notable applications are:

1. **Webteb:** The application launched in 2011, aiming at providing comprehensive health-related information in Arabic platform publishes evidence-based medical information, provides licensed content from established global organizations and academic institutions. The platform provides decision-support tools including WebTeb's Symptom Checker, Drugs & Treatments, Vitamins, Examinations, Vaccinations, Questions, and Answers, among other things.
2. **Altibbi:** Is a digital health platform in the Middle East and North Africa (MENA). Launched in 2008. The platform aims at presenting reliable, up-to-date, and simplified medical information to users in the region in Arabic, according to their proclaimed

¹¹² [Freebase \(database\) - Wikipedia](#)

¹¹³ <https://www.Webteb.com/drug>

¹¹⁴ <https://altibbi.com/>

¹¹⁵ <https://www.123esaaf.com/>

¹¹⁶ <http://www.kuwaitpharmacy.com/Default.aspx>

¹¹⁷ [epharmapedia - \(المراجع الدوائي السوري\) | دليل الأدوية السورية \(المراجع الدوائي السوري\) | Free Medical Apps for Android | أفضل التطبيقات الطبية لأندرويد \(medroid.me\)](#)

¹¹⁸ <http://www.dawee.com>

¹¹⁹ <https://www.adwyaa.com/en>

mission. Within the platform, there exists a section dedicated to medicine where users can navigate through drugs and medicines availability and related information.

3. **123esaaf:** It is an online medical encyclopedia with the most comprehensive and interactive medical resource available online in the Arabic language. It provides credible information, supportive communities, and in-depth reference material about health subjects that matter to all Arab people who search for a certified free source for original and timely health information as well as material from well-known content providers.
4. **Kuwait Pharmacy KP:** Kuwait Pharmacy Information Center is a website established in March 2001. It provides the latest information and news about Medicines among other services.
5. **Epharmapedia:** Syrian Medicines Guide (Syrian Drug Reference): is a guide dedicated to doctors and pharmacists in particular, and in general to ordinary users. It enables search for any medicine by trade name or scientific name form. It is an attempt to achieve the status of the Syrian pharmaceutical encyclopedia among the rest of the pharmaceutical encyclopedias by providing them in the form of an application.
6. **Adwyya.com:** An application established in 2019 in Egypt aiming to enhance the users with all the important information related to medicines such as uses, doses, contradictions, precautions, prices, and drug interactions.

The volume of open data on the Web globally increases rapidly in drug and medicine that opening new opportunities and horizons for enhancing and integrating drug knowledge on a global scale

4.3.2 Linked Open Drug Data LODD

Globally, there exist many Websites that provide drug-related information such as phactMI¹²⁰ (Pharma Collaboration for Transparent Medical Information), Drug Information Portal¹²¹, A to Z Drug Index¹²², and WebMD¹²³. Researchers reported that almost 59% of US adults searched for health information online. However, the number increased to 75% more recently, with more than a billion health-related searches occurring on Google search engines daily. It is more evident that ordinary people's interest in obtaining drug information has increased especially after the new Epidemics such as COVID-19. There is no doubt that individuals are relying more and more on internet search engines regarding their health-related queries and comparisons [336]. The pharmaceutical and drug industry was ahead of other domains in expressing interest in validating the approach for publishing and integrating open data. Several efforts have been made worldwide so far for transforming healthcare and drug data into Linked Data technology. The most notable are the Linking Open Drug Data

¹²⁰ [phactMI](#)

¹²¹ [Drug Information Portal - U.S. National Library of Medicine - Quick Access to Quality Drug Information \(nih.gov\)](#)

¹²² [A - Z Drug List from Drugs.com](#)

¹²³ [WebMD Drugs & Medications - Medical information on prescription drugs, vitamins and over-the-counter medicines](#)

(LODD) project, LinkedCT, Open Biological and Biomedical Ontologies (OBO¹²⁴), and the Semantic Web Health Care and Life Sciences Interest Group (HCLG IG) at W3C.

LODD endpoint was created in 2011, <https://www.w3.org/wiki/HCLSIG/LODD> (courtesy of Anja Jentzsch) [337], which is a pilot study that eases the integration of drug-related data by interlinking and publishing them in the Web of Data and examines use-cases to validate how life science researchers, as well as physicians and patients, can benefit from this Web of Data. Figure 26 shows the incorporation of the datasets published by LODD into the Linked Data cloud. Light grey represents other Linked Data from the life sciences, while white indicates datasets of various domains [337]. LODD project Participants published twelve open-access datasets relevant to pharmaceutical and drug research and development available as Linked Datasets. The published datasets are DrugBank¹²⁵, ClinicalTrials.gov¹²⁶/LinkedCT¹²⁷, DailyMed¹²⁸, SIDER¹²⁹, RxNorm¹³⁰, ChEMBL¹³¹, DisEasome¹³², TCMGeneDIT¹³³/ RDF-TCM, Unified Medical Language System (UMLS)[338], STITCH [339], and Medicare.

LODD Existing tools such as D2R and IBM DB2 have been utilized to present these legacy data as RDF [340]. Each data source is hosted in a discrete store and accessible via a separate SPARQL endpoint. Links between the datasets or to external data sources such as DBpedia are achieved by utilizing software tools capable of automatically creating links between data at a large scale such as Silk [341] and LinQuer [342].

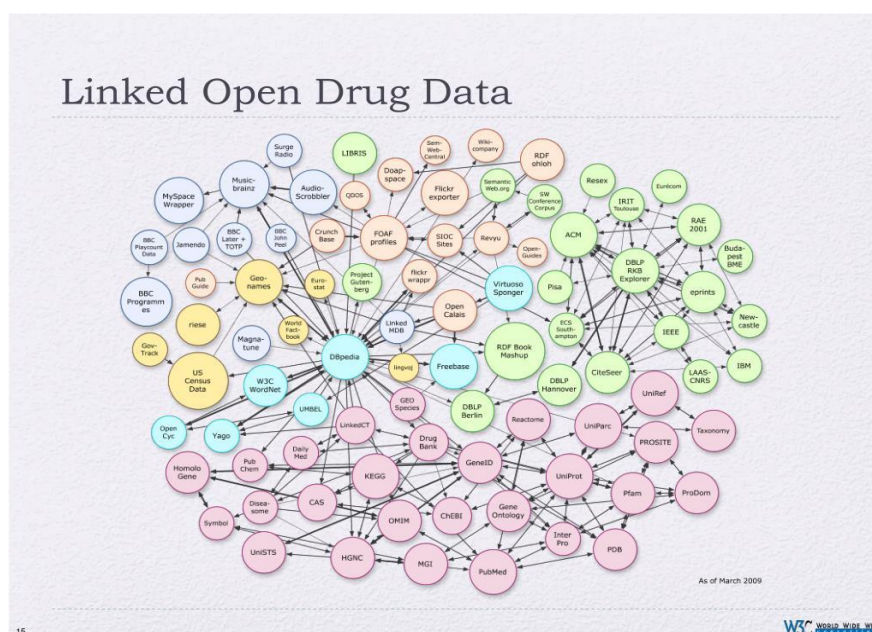


Figure 26: A diagram of the LODD datasets [337]

¹²⁴ [The OBO Foundry](http://www.obofoundry.org/)

¹²⁵ <https://www.drugbank.ca/>

¹²⁶ <https://clinicaltrials.gov/>

¹²⁷ <http://linkedct.org>

¹²⁸ www.dailymed.org

¹²⁹ <http://sideeffects.embl.de/>

¹³⁰ <https://www.nlm.nih.gov/research/umls/rxnorm/>

¹³¹ [ChEMBL Database \(ebi.ac.uk\)](http://ChEMBLDatabase.ebi.ac.uk)

¹³² [DisEasome: an approach to understanding gene-disease interactions - PubMed \(nih.gov\)](http://DisEasome)

¹³³ [TCMGeneDIT: a database for associated traditional Chinese medicine, gene and disease information using text mining - PubMed \(nih.gov\)](http://TCMGeneDIT)

The innovative datasets are intermittently retrieved and the Linked Data representations are refreshed regularly, also, the URIs for representing entities in the linked datasets are unchanging and are chosen by the LODD publishers. As we are investigating open linked datasets, it is worth mentioning that not all of these datasets are considered fully 'open' as outlined by the Panton Principles¹³⁴ (e.g. some of the sources have non-commercial clauses in the license agreement). The LODD project is vigorously exploring the precise situations for alteration and redistribution defined by the data providers and recognizes the boundaries regarding openness. The three most recent additions are RxNorm¹³⁵, Unified Medical Language System (UMLS)¹³⁶, and the WHO Global Health Observatory (GHO)¹³⁷. More details can be found in [343], and a detailed comparison of the LODD datasets can be accessed at <https://www.w3.org/wiki/HCLSIG/LODD/Data> (courtesy of Anja Jentzsch) [337], notably, this page was last updated on 28th December 2012. Later, in 2014, the 3rd release of Bio2RDF (<http://bio2rdf.org/> or <https://github.com/bio2rdf/>) was published as the largest network of Linked Data for the Life Sciences (35 datasets). In 2016, the Linked Drugs (<http://drugs.linkeddata.finki.ukim.mk/>), a dataset was created, which consolidates drug data from 23 countries [344].

LODD has surveyed openly available data about drugs, created Linked Data representations of the datasets, and identified interesting scientific and business questions that can be answered once the data sets are connected. The task force provides recommendations for the best practices of exposing data in a Linked Data representation. LODD datasets established links with datasets provided by other Linked Data projects, such as Bio2RDF [345] and Chem2Bio2RDF [346], as well as primary data providers that offer their resources in RDF, such as UniProt¹³⁸ and the Allen Brain Atlas¹³⁹.

4.4 Interlinking and enhancing Arabic Drugs Datasets

Due to the limitations and lack of Arabic content mentioned above, and the necessity to enhance it and consolidate with additional datasets to meet user requirements. In this thesis, we propose and introduce a solution that will enable Arabic-speaking end-users in general and those who are interested in the drug domain, in particular, to benefit from the Semantic Web technology especially utilizing available linked open data technologies to enrich their datasets. For our use case, we propose to interlink and enhance existing private drug datasets originated in some Arabic countries (see Table 13), with public data and local data enriched with drug information such as Drugbank and DBpedia datasets in the LOD Cloud, see Table 14.

¹³⁴ Panton Principles. <http://pantonprinciples.org/>

¹³⁵ <https://www.nlm.nih.gov/research/umls/rxnorm/index.html>

¹³⁶ [The Unified Medical Language System \(UMLS\) \(nih.gov\)](http://www.nlm.nih.gov/umls/)

¹³⁷ [WHO | Global Health Observatory \(GHO\) data](http://www.who.int/gbo/)

¹³⁸ UniProt. <http://www.uniprot.org/>

¹³⁹ Allen Brain Atlas: Home. <http://www.brain-map.org/>

Table 12: Selected Arabic open drug datasets

Country	Data Set URI	Tuples count	Columns count
Iraq	http://www.iraqipharm.com/upfiles/drug/dreg.xls	9090	9
Lebanon	https://moph.gov.lb/userfiles/files/HealthCareSystem/.../7.../WebMarketed20170307.xls	5822	15
Saudi Arabia	https://www.sfda.gov.sa/en/drug/search/pages/default.aspx	6386	10
Syria	http://www.moh.gov.sy/LinkClick.aspx	9375	7

Table 13: Selected LODD Datasets to be interlinked with Arabic Drug datasets

Dataset	Description
DrugBank	<ul style="list-style-type: none"> • First released in 2006. • A Web-enabled database containing comprehensive molecular information about Chemical, pharmacological and pharmaceutical drug data; data about drug targets (e.g., sequences, structure, pathways). • Contains more than 14575 drugs and 5441 enzyme sequences. • Develop over the years in response to marked improvements to Web standards and changing needs for drug research and development. • Website: http://www.drugbank.ca/.
DBpedia	<ul style="list-style-type: none"> • An ongoing project designed to extract structured data from Wikipedia. • Containing more than 228 million entities to date. • Contains RDF data, about 2.49 million things out of which is 218 million triples describing 2300 drugs. • Updated every three months. • Website: http://www.dbpedia.org/.

Even if the datasets were created in Arabic speaking countries, they are prepared in the English language, as it is the most widely used language by doctors and pharmacists in the Arab countries (Tunisia, Algeria, and Morocco; French Language is the dominant scientific language), but the ordinary people are less acquainted with the English language. Also, these datasets contain a few columns which make them lack the most needed information by users such as side-effects, abstract in Arabic language, similar drugs, usage in certain cases such as in pregnancy, and prices, etc.

Our goal of the innovative Arabic drug application proposed in this thesis is to enable end-users to pose inquiries about drug availability in the open datasets (e.g., DrugBank, DBpedia, see Table 14 above) and to enrich the local data store with information from the LOD Cloud, i.e., integrating public and private datasets as in Figure 1. The end-user will profit from the interlinking of private datasets with open data and improvement of local data with information from the Web. Examples of key business queries include but are not limited to:

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 selected parts of textual documentation of the product, specifically extracted from Summary of Product Characteristics.

3. For a particular drug, retrieve equivalent drugs, and compare their active ingredients, contradictions, and prices.
4. For a particular drug, retrieve valuable information about equivalent drugs with different commercial names, manufacturers, strengths, forms, prices, etc.
5. For an active ingredient show advanced clinical information i.e., pharmacological action, pharmacokinetics etc.
6. 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, chronic diseases).
7. For a particular active ingredient, retrieve advanced clinical information, i.e., pharmacological action, pharmacokinetics, etc.
8. Retrieve interactions in a set of medicinal products and/or active ingredients.
9. For a particular drug, retrieve its cost, manufacturer, and country.

4.4.1 Selection of Linked Open Data Tools

There exists an array of tools that are available for linked data and linked open data to support their projects and activities such as creation and conversion of RDF serialization, metadata mapping and editing, data modelling, storage and access components, searching, discovery and publishing. The most used tools for linked data are *OpenRefine*¹⁴⁰, *Virtuoso Sponger*¹⁴¹, *RDF Mapping Language*¹⁴², *RDF123*¹⁴³, *Karma*, *XLWrap*¹⁴⁴, *csv2rdf4lod*¹⁴⁵, *Tarql*¹⁴⁶, *TopBraid Composer*¹⁴⁷, *TabLinker*¹⁴⁸, *D2R Server*¹⁴⁹, *Silk Framework*¹⁵⁰, *Triplify*¹⁵¹, and *D2RQ*¹⁵². Table 15 presents a list of some of the available tools used for the RDF transformation process.

Table 14: RDF Transformation tools

Tool	Description
Virtuoso Sponger	<ul style="list-style-type: none"> • Is Virtuoso's middleware for generating Linked Data from a variety of data and formats, transparently integrated into Virtuoso's SPARQL query processor. • The main functionality is provided by Cartridges. Each cartridge includes Data Extractors to extract data from one or more data sources, and Ontology Mappers to annotate the extracted data to a certain schema to generate Linked Data.

¹⁴⁰ www.openrefine.org

¹⁴¹ vos.openlinksw.com/owiki/wiki/VOS/VirtSponger

¹⁴² <https://github.com/RMLio>

¹⁴³ [RDF123 - Mathematical software - swMATH](http://rdf123.sourceforge.net/)

¹⁴⁴ <http://xlwrap.sourceforge.net/>

¹⁴⁵ <https://github.com/timrdf/csv2rdf4lod-automation/wiki>

¹⁴⁶ [Tarql: SPARQL for Tables – Tarql – SPARQL for Tables: Turn CSV into RDF using SPARQL syntax](http://tarql.org/)

¹⁴⁷ <https://www.topquadrant.com/tools/IDE-topbraid-composer-maestro-edition/>

¹⁴⁸ <https://github.com/Data2Semantics/TabLinker>

¹⁴⁹ [D2R Server: Accessing databases with SPARQL and as Linked Data. http://d2rq.org/d2r-server](http://d2rq.org/d2r-server)

¹⁵⁰ [Silk Framework. http://silkframework.org/](http://silkframework.org/)

¹⁵¹ [Triplify - Semantic Web Standards \(w3.org\)](http://triplify.org/)

¹⁵² [The D2RQ Platform – Accessing Relational Databases as Virtual RDF Graphs](http://d2rq.org/)

	<ul style="list-style-type: none"> • Relies on custom scripts for each different format to generate the corresponding Linked Data.
OpenRefine	<ul style="list-style-type: none"> • A tool from Google, previously known as Google-Refine used to explore, clean, reconcile and transform messy uncleaned datasets within organizations. • Transformation (normalizing and de-normalizing functions) using Google Refine Expression Language (GREL) and exports the refined data in TSV, CSV, Excel and HTML table formats. • Permits input is in the format CSV, XML, JSON, RDF triples, spreadsheet, • Helpful in the cleaning process for removing data inconsistencies within the raw data. • The schema mapping can be defined in a graphical UI. • Reconciles against SPARQL endpoints, RDF dumps and searches the Web for related Linked Data sets. • Allows users to define how the raw data is modelled as Linked Data by importing their vocabularies or reusing existing ones through its user interface. • Rules that human agents define can only be exported in a custom JSON format and their execution is only performed by Open Refine via its custom scripts.
RDF Mapping Language	<ul style="list-style-type: none"> • Used for specifying customized mappings from heterogeneous data structures and serializations (including databases, XML, CSV) to the RDF data model.
RDF123	<ul style="list-style-type: none"> • Highly flexible open-source tool. • Used to transform spreadsheet data to RDF. • Adaptable to Windows and Linux applications to download, a Java application and servlet.
Karma	<ul style="list-style-type: none"> • Enables users to integrate data from a variety of data formats, such as data in databases, spreadsheets, XML, JSON, and KML. • Users can Semantically annotate their data according to a Semantic schema of their choice, relying on Karma's user interface that automates much of the process. • Turns all different data formats into a tabular structure, following the rely upon Nested Relational Model (NRM) as an intermediate form to represent data.
XLWrap	<ul style="list-style-type: none"> • Wraps spreadsheets (including cross tables) to arbitrary RDF graphs. • Supports Excel/Open Document/CSV streamed processing, local/HTTP loading, expressions similar to Excel/OpenOffice Calc, custom functions, usage via API or SPARQL endpoint.
csv2rdf4lod	<ul style="list-style-type: none"> • Uses declarative RDF enhancement parameters to specify how to transform tabular data into well-structured and well-connected RDF. • Uses identifiers for source organization, dataset, and version to establish default namespaces for all URIs created and provides VOID and provenance metadata as part of the conversion output.
Tarql	<ul style="list-style-type: none"> • A command-line application that converts CSV to RDF with a user-defined mapping written in SPARQL 1.1 (standard).
TopBraid Composer	<ul style="list-style-type: none"> • Converts Excel spreadsheets into instances of an RDF schema.
TabLinker	<ul style="list-style-type: none"> • Converts non-standard Excel spreadsheets to the Data Cube vocabulary, e.g., Excel files that contain hierarchical information in row and column headers etc.

Sheet2RDF	<ul style="list-style-type: none"> • A platform for the acquisition and transformation of spreadsheets into RDF. • Combines a practical user interface with the potentialities of a full transformation PEARL.
Sparqlify	<ul style="list-style-type: none"> • Is a SPARQL-SQL rewriter that enables to define RDF views on relational databases and query them with SPARQ.
Spread2RDF	<ul style="list-style-type: none"> • A converter for complex spreadsheets to RDF and a Ruby-internal DSL for specifying the mapping rules for this conversion.
Triplify	<ul style="list-style-type: none"> • A PHP plugin reveals the Semantic structures encoded in relational databases by making database content available as (RDF, JSON, or Linked Data).
D2RQ	<ul style="list-style-type: none"> • A Platform is a system for accessing relational databases as virtual and read-only RDF graphs, also offers RDF-based access to the contents of the relational database without replicating it into an RDF store.

4.5 ALDDA Piloting methodology and QA framework development

After reviewing previous attempts to implement linked data applications discussed in previous sections, in this thesis, we are implementing a novel methodology for linked data and we propose to split the implementation of a linked data application development into three main software development phases based on our previous work [196]. The process of developing a new pilot application [196] using open-source linked data tools (e.g., from the linked data stack) can be divided into three phases: (1) initialization; (2) innovation; and (3) validation; as presented in Figure 27 and Table 16.

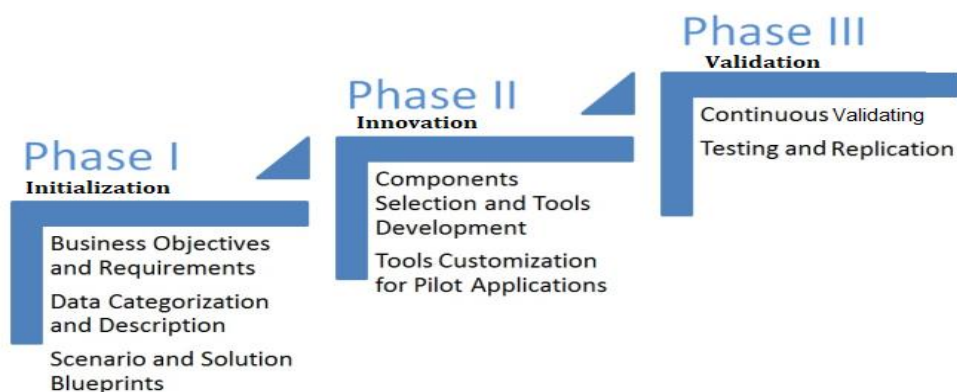


Figure 27: Piloting methodology phases

Table 15: Linked data application phases and detailed steps

PHASE	DESCRIPTION
INITIALIZATION	<ul style="list-style-type: none"> • The standard models (vocabularies, taxonomies) are selected for structuring and describing the data. • scheduled data extraction and loading operations. • Test linked data components are (open-source tools) for processing and/or data transformation.

This Phase mainly handles the data preparation operations and constitutes the data selection, data analysis, and data cleaning processes. The quality assessment at this phase guarantees appropriate selection and analysis of datasets and applying acceptable data cleaning steps.

In particular, the initialization phase will perform:

1. 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);

2. Data categorization and description: Analysis of the datasets to be published in linked data format and selection of vocabularies and development of other specifications for metadata description;

3. Scenario and Solution Blueprints:

- Comprehensive scenarios to showcase the power of novel application
- Analysis of functionalities
- Requirement's specification document

Example: In addition to corporate data, the targeted data is selected from the Arabic drug datasets mentioned above) along with the public datasets (DrugBank and DBpedia). Appropriate vocabularies are selected or developed, and mapping rules are defined.

Selected linked data components (open-source tools) are customized and developed to match the needs of the target application. This phase usually includes integration with existing enterprise systems and the adoption of proven technologies for the benefit of the end-user organization.

This stage constitutes the *Ontology definition, Schema mapping, data modeling, creating an RDF dataset, Data Interlinking, Data publishing, quality assessment, and data storage*. Again data quality assessment procedures are implemented at every step (revise ontologies, test data modelling, etc.) to ensure an acceptable level of quality before moving to the next phase

In particular, the stage will perform:

1. 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, modelling, and conversion.

The fundamentals of ontology engineering and utilization have been established to allow better alignment between datasets through maximizing the chances of reuse [11]. In another word, the data

INNOVATION

	<p>publisher should always attempt to reuse an existing vocabulary or ontology, giving priority to the most used. Several tools for ontology and vocabulary discovery exist that data publishers are encouraged to use in this stage. The two most prominent are Linked Open Vocabularies (LOV¹⁵³) and DERI¹⁵⁴ Vocabularies, which are also used to provide usage statistics to assess the effect of a given vocabulary or ontology in a specific domain.</p> <p>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.</p> <p>2. Generic component selection and tool customization for the pilot applications: Customization of linked data components for use in the targeted domain.</p> <p>Example: In this phase, tools for federated search and data are selected. Additionally, big data analytics tools are selected, custom visualization and user interfaces are created [347].</p>
<p>VALIDATION AND SPECIFIC TOOLS DEVELOPMENT</p>	<p>Towards the end of the development, data quality is performed on the final data outcome through qualified testers who understand the business requirements collaborate with engineering until fully operable and enterprise-ready tools are on market.</p> <p>In this phase open-source tools are validated for reuse; feedback is provided for improving the solution components and ensuring quality, and new interfaces are built.</p> <p>1. Continuous Validation</p> <ul style="list-style-type: none"> ○ Quality assessment; ○ Tool for Workflow Automation; ○ Open-source tools are validated for reuse; ○ Feedback is provided for improving the solution components; and <p>2. Testing and Replication:</p> <ul style="list-style-type: none"> ○ Business users ○ Citizens

4.6 Selection and Implementation of Data Quality Assessment Measures

Data quality assessment is a good starting point to identify insignificant information in datasets. A preliminary discussion on Data Quality assessment challenges and possible solutions for big data scenarios has been proposed by Cappiello *et al.* (2018) stated that “Whenever the data source is updated the data quality assessment also changes” [142].

¹⁵³ Linked Open Vocabularies (LOV). <http://lov.okfn.org/>. Accessed: 01-12-2020.

¹⁵⁴ DERI Vocabularies. <http://vocab.deri.ie/>. Accessed: 01-12-2020

Data quality may be measured *subjectively*, by asking data consumers to assess the quality level of the dimensions. Instead, data quality metrics may be defined to measure dimensions of data quality *objectively*. Data quality assessment must also deal with subjective perceptions of the individuals dealing with the data, it reflects the needs and experiences of data collectors, trustees, and consumers of data products [294][295]. Subjective assessments examine a stakeholder's subjective perception about data quality, typically in a questionnaire form. Objective measurements based upon the concerned dataset can be task-independent or task-dependent. If objective measures cannot be applied for some data quality dimensions assessment, , subjective measures are applied instead [296].

Pipino et al. (2002), addressed 16 subjective and objective common types of data quality dimensions, ordered in alphabetical order: *accessibility, an appropriate amount of data, believability, completeness, concise representation, consistent representation, ease of manipulation, free-of-error, interpretability, objectivity, relevancy, reputation, security, timeliness, understandability, and value-added* [265]. *Heinrich et al. (2011)*, presented 6 requirements needed for the data quality measures process: *normalization, interval scale, interpretability, aggregation, adaptability, and feasibility*. The authors presented metrics for *correctness* and *timeliness* that encounter these requirements. The *correctness* measure calculates the percentage of the distances between the data and real outcomes, whereas, the *timeliness* measure is defined by an exponentially decaying function in terms of decline and age [297].

Usually, one metric is insufficient to measure data quality dimension accurately, instead of combining different metrics to have a better view of the inclusive data quality. Some metrics measure the percentage of the number of specified constraints that are being violated or count the number of erroneous decisions made based on the data [298]. For example, consistency, which is a form of reliability, can be measured by Cronbach's alpha [299]. In literature, there are three forms of objective assessment "*Simple Ratio*", "*Min or Max Operation*", and "*Weighted Average*" [265], Table 10 below shows a brief comparison of the above functional forms of data quality assessment.

Table 16: Data quality assessments functional forms

Functional Form	Description	Dimensioned measured
<i>Simple Ratio</i>	The ratio of the desired outcomes to the total outcomes	Accuracy, completeness, consistency, conciseness, relevancy, ease of manipulation
<i>Min/Max Operations</i>	The minimum or maximum value among normalized individual data quality indicator values	Believability, an appropriate amount of data, timeliness, and accessibility
<i>Weighted Average</i>	Assigning weighting factors to represent the importance of the variables to the evaluation of a dimension	Believability, an appropriate amount of data

○ **Simple Ratio:** measures the ratio of desired outcomes to total outcomes, i.e., it measures the functional proportion of valid records out of total records [265]. This simple ratio committed to the convention that 1 represents the most desirable and 0 the least desirable score [300][294][227][248]. Many traditional data quality metrics, such as *completeness*, *accuracy*, and *consistency* take this form. *Concise representation*, *relevancy*, and *ease of use* dimensions can be evaluated utilizing simple ratio functional form as follows:

- a. **Accuracy** dimension, If the data units in error are counted, the metric can be defined as the number of data units in error divided by the total number of data units subtracted from 1, and have the following form:

$$Accuracy = 1 - \left(\frac{Ncv}{N}\right); \text{ where } Ncv = \text{Number of Correct Values and;} \\ N = \text{Total Number of values of the sample dataset}$$

- b. **Consistency** dimension can be viewed from different perspectives, such as; the consistency of redundant data values in one or multiple tables, or the consistency between two related data elements, or the consistency of format for the same data element used in different tables.

The metric measuring consistency dimension is the ratio of violations of a precise consistency type to the total number of consistency checks subtracted from 1 [265]. The metric to measure consistency may have the following form:

$$Consistency = 1 - \left(\frac{Ncnv}{N}\right); \text{ where; } Ncnv = \text{Number of consistent values} \\ N = \text{Total Number of values of the sample dataset}$$

- c. **Relevancy** dimension, there is no obvious information measurement in the literature. Relevancy may be measured as a ratio between the number of the relevant keywords in the description field to the total number of words in the specific domain in a dataset.

$$Relevancy = 1 - \left(\frac{Nrkd}{N}\right); \text{ where;} \\ Nrkd = \text{Number of relevant keywords in the description} \\ N = \text{Total Number of words in the sample dataset}$$

For the above definitions and metrics, the value of accuracy, consistency, and relevancy ranges between 0 and 1.

○ **Min or Max Operation:** Used to measure dimensions that necessitate the aggregation of multiple data quality indicators (variables), where the minimum or maximum operation can be applied. The minimum (or maximum) value is computed from the normalized values of the individual data quality indicators. The min operator is conventional in that it assigns to the dimension an aggregate value no higher than the value of its weakest data quality indicator (evaluated and normalized to between 0 and 1)[213]. The maximum operation is utilized if a liberal interpretation is warranted [264]. The individual variables are measured by a simple ratio [213]. This function can be used for the compute dimensions *believability*, *an appropriate amount of data*, *timeliness*, and *accessibility*.

○ **Weighted Average:** It is a calculation that takes into account the differing degrees of importance of the numbers in a dataset¹⁵⁵. In calculating a weighted average, every number in the data set is multiplied by a predetermined (coefficient) weight before the final calculation is performed. It is regarded as an alternative to the min operator of variables where organizations and companies comprehend the importance of all variables to the overall evaluation of a particular dimension, then the calculation of the weighted average of the variables is convenient. The metric to measure consistency may have the following form:

$$\text{Weighting Average} = \sum_{i=1}^n A_i M_i \text{ Where } A_i: \text{ is the weighting coefficient, } 0 \leq A_i \leq 1, \\ \text{and } A_1 + A_2 + \dots + A_n = 1.$$

M_i : is a normalized value of the assessments of the i^{th} variable.

To ensure the rating is normalized, each weighting factor should be between zero and one, and the weighting factors should add to one.

4.7 Validation of the ALDDA Approach

In this thesis, we extend the above methodology to meet the requirements for the Arabic Linked Drug Dataset application (ALDDA) development, which is mainly focused on gaining additional valued information from open data to enhance the drug datasets, consolidate in a form of a Semantic Data Lake. The methodology consists of five main phases (*Initialization Phase; Innovation Phase; Conversion phase; Quality assessment Phase; and Visualization and Querying Phase*) as presented in Figure 28:

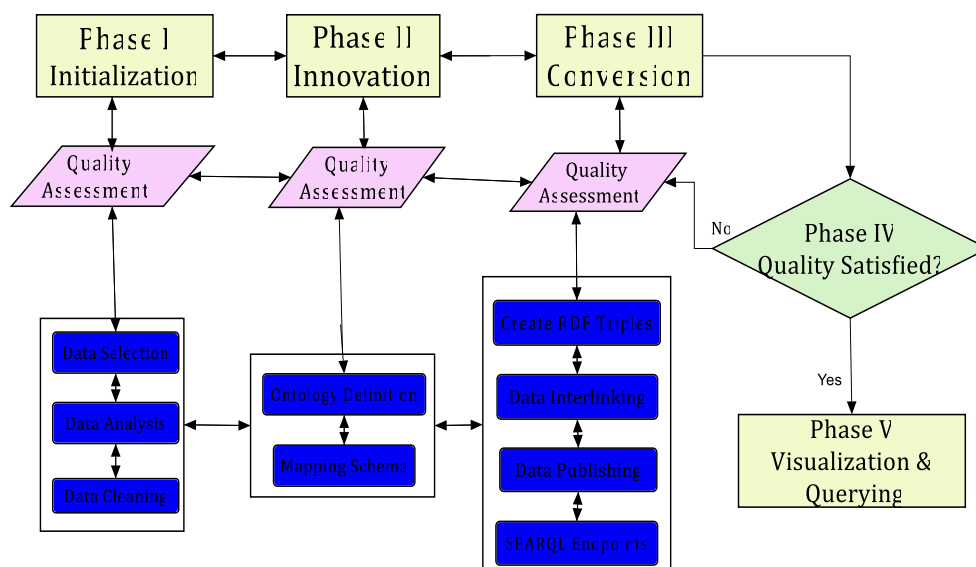


Figure 28: A novel linked data methodology with a focus on quality assessment

As mentioned, many times earlier that our case study uses datasets from four different organizations in four different countries, which strongly suggests that there will be major

¹⁵⁵ Weighted Average Definition (investope dia.com)

inconsistencies as there are no prior agreed guidelines to control the design of the data files between the four organizations in those countries. For this reason, it is suggested that in our methodology quality assessment should be carried out at every phase of the methodology and an overall quality assessment check after completing the three top phases to guarantee as good combined dataset quality as possible. In what follows, we will discuss the methodology stages.

4.7.1 Implementation of the Arabic Linked Drug Data Application (ALDDA)

As illustrated in previous sections of this chapter, the selected data were static inputs i.e XLS Excel spreadsheets. To use them as Linked Data on the Web, they must undergo a conversion process that outputs static RDF files or loads converted data directly into an RDF store. Some of the tools used for RDFization process are listed in Table 16. As far as this thesis is concerned, we selected the OpenRefine tool to perform the conversion process. The implementation of the ALDDA comprises the following phases and steps:

I.Data Initialization Phase (Preparation):

This phase consists of the following three processes:

- **Data Selection:** As a use case scenario, after serving the Web for Arabic drug datasets, four drug data files were selected from four different Arabic countries, namely, Iraq, Saudi Arabia, Syria, and Lebanon (see Table 14, section 4.4). Most of the open published files in the Arab region are either in PDF, XLS, JPG format. The reasons for choosing the XLS format were data fidelity, the ability to source from a wider range of public sector domains, and to have increased value that comes from many information linkages. We 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 every two years. 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.
- **Data Analysis:** After analyzing the data quality of the selected files, it appears that the overall quality is rather 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 are in a form of a generic online drug database, see Table 13. These two databases contain 13,445 records. To gather the data in HTML format, we built HTML Crawlers based on JSOUP¹⁵⁶, 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 we had to use their XLS files and implement additional transformations to extract active ingredient information.
- **Data Cleaning:** OpenRefine (version 2.6-rc1) was used to clean the selected data to make it coherent and ready for further operations according to the methodology. This procedure permitted us to create links between the datasets in the course of transformation into RDF. A well-organized cleaning operation minimizes inconsistencies and ensures data

¹⁵⁶ <http://jsoup.org>

standardization among a variety of data sources. Data cleaning aligns and transforms the XLS initial data into 5-star Linked Data and publishes them on the Web in a common, aligned, and combined Linked Drug Data dataset. Raw institutional datasets occasionally contain several inconsistencies and may lack the standard representation format. As expected, our selected datasets contain several inconsistencies and lack a standard representation format. It should be noted that the selected data does not have good quality standards, due to many reasons discussed above, so a lot of analysis and cleaning is required, for this reason, quality assessment is assigned on top of this phase as well the other two phases. In another word, the quality assessment is an ongoing process throughout the whole methodology to ensure acceptable output to the end-user.

II. Modeling and Innovation Phase (Integrating Datasets in the form of a knowledge graph):

i) Data Modeling (Mapping Schema and Ontology Definition):

• Data Mapping

It is necessary to define the mappings between new classes and properties and the classes and properties from other ontologies each time a new ontology is developed to enable ontology matching and RDF-based reasoning, for schema alignment. The data mapping process involved importing the CSV files (created from the selected XLS files) into relational databases in Virtuoso and using the LODrefine to transform the RDB data into RDF data to produce an RDF graph by using the RDF-extension. The produced RDF files permits utilizing Semantic Web technologies such as SPARQL querying over data that resides in standard relational databases. The naming scheme for the selected file's attributes are different as can be seen in the following tables, the following data mappings were necessary to be carried out for the selected data files as in Figure 29:

1. IRAQ Data file

Original Attribute	Mapped Attribute
Scientific name	genericName
Trade name	brandName
Packaging & dosage form	dosageForm
Authorization holder (manufacturer)	manufacturer1
No. & date of registration	licenceValidFrom

2. Syria Data file

Original Attribute	Mapped Attribute
Scientific name of the preparation	genericName
The commercial name of the product	brandName
Name	Manufacturer1
Caliber	Amount
Package	dosageForm
Price for the public	CostPerUnit

3. Saudi Arabia (Web database) Data file

Original Attribute	Mapped Attribute
Generic Name	genericName
Trade Name	brandName
Strength Value	strengthValue1
DosageForm	dosageForm
Manufacturer Name	manufacturer1
Price	costPerUnit
Registration No	licenceValidFrom
Volume	Amount

4. Lebanon (Web database) Data file

Original Attribute	Mapped Attribute
ATC	atcCode
Ingredients	activeSubstance1/ activeSubstance2/ activeSubstance3/ activeSubstance4/ activeSubstance5/strengthValue1/ strengthValue2/ strengthUnit1/ strengthUnit2
Name	brandname

<i>Dosage</i>	<i>dosageForm</i>
<i>Laboratory</i>	<i>manufacturer1</i>
<i>Price</i>	<i>costPerUnit</i>
<i>Registration No</i>	<i>licenceValidFrom</i>
<i>Exch_date</i>	<i>licenceValidUntil</i>

Figure 29: Data mapping from Arabic datasets

After mapping the data files, we merged them in one CSV file, we called it the ALDDA file, as in Table 17.

Table 17: The ALDDA merged property file after mapping

<i>ALDDA property</i>	<i>Description</i>
<i>dosageForm</i>	Drug dosage form
<i>activeSubstance1</i>	Product active substance
<i>activeSubstance2</i>	Product active substance
<i>activeSubstance3</i>	Product active substance
<i>activeSubstance4</i>	Product active substance
<i>activeSubstance5</i>	Product active substance
<i>strengthValue1</i>	Product concentrations
<i>strengthValue2</i>	Product concentrations
<i>strengthUnit1</i>	Product concentration unit
<i>strengthUnit2</i>	Product concentration unit
<i>Manufacturer1</i>	Product manufacturer name
<i>costPerUnit</i>	Product unit price
<i>licenceValidFrom</i>	License validation starting date
<i>licenceValidUntil</i>	License validation finishing date
<i>Amount</i>	Product available quantity

As can be seen from the selected drug data (Table 17), the published public drug data comprise different sets of information. Each instance of the drug class has properties such as generic drug name, code, active substances, non-proprietary name, strength value, cost per unit, manufacturer, related drug, description, URL, license, etc. Additionally, ATC code is used for cataloging drugs and it is controlled by the World Health Organization. After unifying the original attributes in the previous step, an ontology is needed to transform and represent the drug data in RDF.

The ontology development was based on re-use of classes and properties from existing ontologies and vocabularies including Schema.org vocabulary¹⁵⁷, DBpedia Ontology¹⁵⁸, UMBEL (Upper Mapping and Binding Exchange Layer)¹⁵⁹, DICOM (Digital Imaging and Communications in Medicine)¹⁶⁰, and DrugBank in addition to other biomedical ontologies as they cover the properties we needed and provide us easier interlinking possibilities for additional transformation. Obeying the best practices for ontology development, we decided to re-use existing drug ontologies, the DrugBank RDF repository and its ontology that would enable us

¹⁵⁷ <https://schema.org/>

¹⁵⁸ <https://wiki.dbpedia.org/services-resources/ontology>

¹⁵⁹ <http://umbel.org/>

¹⁶⁰ <https://www.dicomstandard.org/>

for the interlinking process later. Additionally, to align the drug data with generic drugs from DrugBank properties, the following DrugBank properties were used:

brandName	(The brand name of the drug);
genericName	(The generic name of the drug);
atcCode	(The global ATC code of the drug); and
dosageForm	(the pharmaceutical form of the drug)

The 'drugs' class contained in the DrugBank ontology represents the drug entities, as well as the relations for the ATC code, the generic name, brand name, and the dosage form. The 'drugs' class along with the 'atcCode', 'genericName', 'brandName', and the drugDosageForm properties were used in our ontology. Additional drug information is required but not covered by the DrugBank ontology, such as unit price. Hence, we developed our ontology: the ALDDA ontology. Our ALDDA ontology comprises a class for drug type entities, named 'ADrug' (Figure 30).

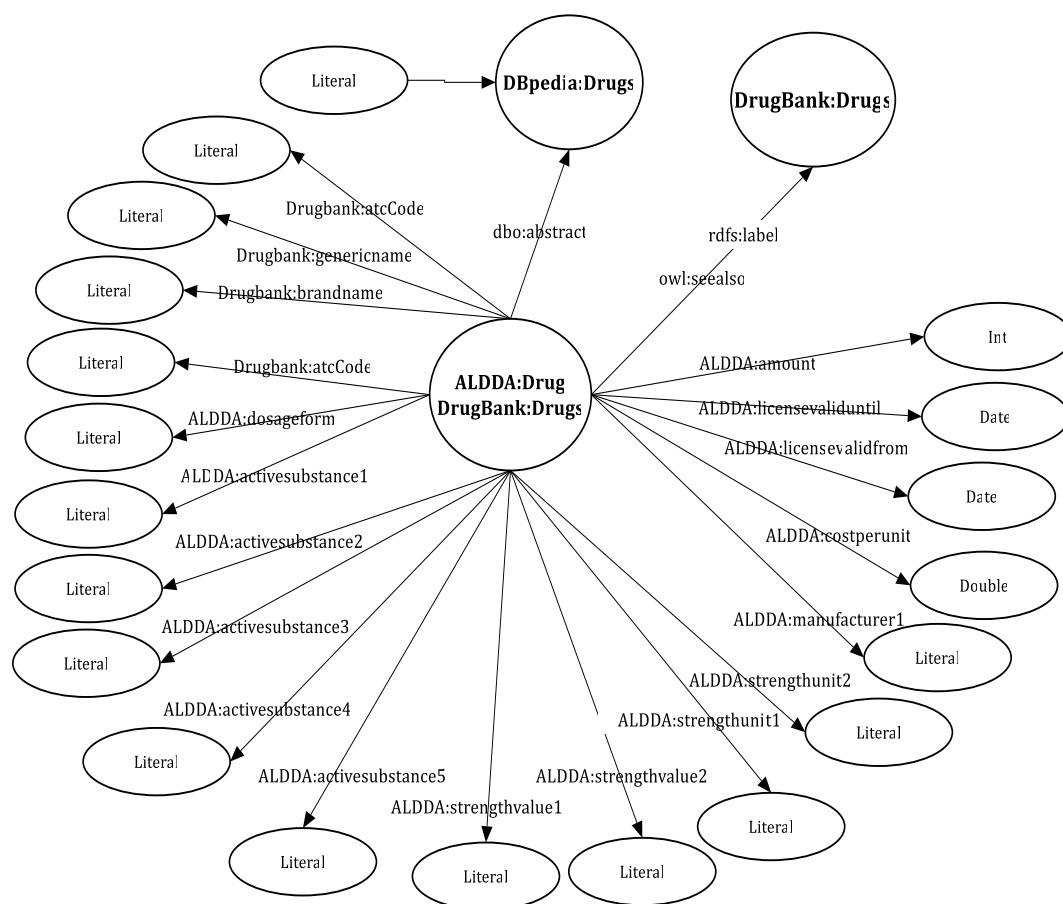


Figure 30: The ALDDA: Drug class

In addition to the properties used from DrugBank and the ALDDA properties defined in the ontology, we use the 'rdfs:label' and 'owl:seeAlso' properties. The 'rdfs:label' property is used to point to the generic name of the drug, whereas 'owl:seeAlso' is used to link the drugs from our HIFM graph with drugs from DrugBank. 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)¹⁶¹ and DERI Vocabularies¹⁶².

III. Data Conversion Phase (Creating RDF, Interlinking, publishing, and Querying)

This phase of the methodology transforms the cleaned source data files into the RDF schema. It contains the necessary steps of the conversion process as follows:

1. Create RDF dataset:

This step transforms raw data into an RDF dataset based on a serialization format. The actual transformation process can be encapsulated in an automated script that gets the source dataset which conforms to the template, sends it to the transformation tool, and gets the outputted RDF. To achieve the transformation of our drug data into 5-star LOD, we are required to have relations in the RDF graph towards external entities. Therefore, we choose to use the DrugBank dataset, as it comprises the most detailed drug dataset on the cloud. Likewise, we needed relations in the next step, the interlinking process, so we used the ATC codes from Drugbank to perceive the similarity between our dataset drugs and Drugbank drugs.

2. Data Interlinking:

After the drug dataset is transformed into a Linked Data dataset i.e., creating the RDF file in the previous process, we need to create the internal links between drugs that share the same use. To create these links, we use the drug's ATC codes. According to the WHO coding scheme¹⁶³, if two drugs have the same ATC code, they share the same function. Our LODRefine transformation script is designed for data fulfilling the CSV template, and its output is a Linked Drug Data dataset that uses our defined RDF schema. The transformation process starts with reconciling the columns *atcCode*, *genericName1*, *activeSubstance1*, *activeSubstance2*, *activeSubstance3*, *activeSubstance4* and *activeSubstance5* reconciled with *DBpedia*, then creates an RDF schema skeleton. 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 relation 'owl:seeAlso' from the commonly used OWL namespace was preferred over the 'owl:sameAs' relation, since it cannot be guaranteed that the descriptions of two drugs refer to the same real-world entity i.e., drug. As an example, a specific drug in our dataset includes information about a manufacturer, active substances, dosage form, strength, validity, and price. On the other hand, a drug in the DrugBank dataset contains information about the chemical formula, molecular weight, affected organisms, contraindications, interactions, etc., i.e., information about drugs as technical information not as a product for dispensing. The activities

¹⁶¹ <http://lov.okfn.org/>

¹⁶² <http://datahub.io>

¹⁶³ ATC Codes: Structure and Principles. http://www.whocc.no/atc/structure_and_principles.

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 showcases the total number of links generated between the source and target data sources. Several reconciliation services were built to make successful interlinking as in Figure 31.

1. DBpedia Reconciliation service based on atcCode	2. DBpedia Reconciliation service based on genericName
<pre> PREFIX drugbank: <http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugbank/> PREFIX dbo: <http://dbpedia.org/ontology/> SELECT * WHERE { ?s dbo:atcPrefix ?atcPrefix . OPTIONAL { ?s dbo:atcSuffix ?atcSuffix . } BIND (concat(?atcPrefix, ?atcSuffix) AS ?atcCode) FILTER regex(?atcCode, '<drugAtcCode>') } </pre>	<pre> PREFIX drugbank: <http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugbank/> PREFIX dbo: <http://dbpedia.org/ontology/> SELECT str(?label) ?s WHERE { ?s rdf:type dbo:Drug . ?s rdfs:label ?label . FILTER regex(?label, '<drugGenericName>') } </pre>
3. DBpedia reconciliation service which filters entities of type ChemicalSubstance	4. DBpedia reconciliation service which retrieves genericName in Drug synonyms
<pre> PREFIX drugbank: <http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugbank/> PREFIX dbo: <http://dbpedia.org/ontology/> SELECT str(?label) ?s WHERE { ?s rdf:type dbo:ChemicalSubstance . ?s rdfs:label ?label . FILTER regex(?label, '<drugActiveSubstance>') } </pre>	<pre> PREFIX drugbank: <http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugbank/> PREFIX dbo: <http://dbpedia.org/ontology/> SELECT ?label ?s WHERE { ?s rdf:type dbo:Drug . ?s rdfs:label ?label . ?s dbp:synonyms ?synonyms FILTER regex(?synonyms, "<drugGenericName>") } </pre>

Figure 31: The reconciliation process based on actCode, genericName, ChemicalSubstances, and Drug synonyms

The URI of the first result of the sequential execution of previous services is used to make additional `rdfs:seeAlso` attribute which is used in interlinking with DBpedia.

3. Data Publishing and Storage:

Publishing data on the web according to the principles of linked data enables data providers to add their data to global data space, making it discoverable and useable by numerous applications. Publishing a dataset as linked data on the web requires assigning URIs to the entities described by the dataset and provide for dereferencing these URIs over the HTTP protocol into RDF representations, setting RDF links to external data sources on the web so that clients can navigate the web of data as a whole by following RDF links, and providing metadata about published data so that clients can assess the quality of published data. A variety of linked data publishing tools has been established, which either serve the RDF stores content as linked data or provide linked data views over non-RDF inheritance data sources. These tools permit publishers to avert dealing with technical details such as content negotiation, to ensure that data is published according to the linked data community best practices.

After a graph of Linked Open Data has been created from the diverse Arabic drug datasets, we start the next process of data publishing on the Web. Data publishing ought to be carried out according to the W3C recommendations for publishing Linked Open Data on the Web. The recommendations propose enabling direct URI resolution, providing a RESTful API, providing a SPARQL endpoint, and/or providing the dataset as a file for download [321]. OpenLink Virtuoso server (version 06.01.3127)¹⁶⁴ on Linux (x86_64-pc-Linux-gnu), Single Server Edition have been used as a triple store, this public instance of Virtuoso contains the Linked Drug Data from the ALDDA graph, and provides a public interface via its SPARQL endpoint queries on 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/>.

Drug data from the graph can be queried by using the SPARQL editor available at the endpoint, or by using the endpoint as a Web service from the Web, desktop application, or a mobile client. The endpoint can be utilized as a Web service by adding the SPARQL query into a query string, appended to the URL of the endpoint. Virtuoso among other tools and platforms permits Linked Data publishing of datasets created originally in an RDF file (Turtle, N3, RDF/XML, JSON-LD, etc.), a CSV file, or in a relational database.

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 (*REpresentational State Transfer*) has been applied to describe the desired web architecture, to identify existing problems. REST API is used to make a Web application. Rest API can separate the Front-end and Back-end of a website, and it is also a good way for providing web services, so Back-end APIs designed in a Rest style are becoming popular nowadays. 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.

IV. Specific tools development and validation:

1. **Tools for Quality Assessment:** In our approach, quality assessment is an ongoing operation in all phases of the methodology as the quality of the content of the document on the Web varies [348][349]. We 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 data quality 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 data quality assessment suggestions indicating the quality constraints that will increase or

¹⁶⁴ <https://github.com/openlink/virtuoso-opensource>

decrease the data quality. We also believe that data quality 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 data quality 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 data quality are within a range from 0 to 1, with 0 representing an 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 data quality dimensions relevant for quality assessment of Arabic DBpedia, we have identified three dimensions accuracy, consistency, and relevancy, as shown in Table 18.

Table 18: Data Quality dimensions relevant for quality assessment of Arabic DBpedia (*Specific to DBpedia, **Specific to Arabic DBpedia)

Dimension / Metrics Definition	Category	Sub-category
<p>Accuracy (Intrinsic): I Is the degree of closeness between a value x and a value x', considered as the correct representation of the reality that x aims to represent.</p> <p>If x is the number of the correct values, and x' is the number of total values, then, $Accuracy = x/x'$</p>	Triple incorrectly extracted	<ul style="list-style-type: none"> • Object value is incorrectly/ incompletely extracted • Special template not properly recognized* • Wrong values in numerical data**
	Data type problems	<ul style="list-style-type: none"> • Data type incorrectly extracted
	Implicit relationship between attributes	<ul style="list-style-type: none"> • One/ Several fact encoded in one/several attributes* • Attribute value computed from another attribute value**
<p>Consistency (Intrinsic): Data is consistent if it meets a set of constraints. If x is the number of consistent values, and x' is the number of total values. Then, $consistency = x/x'$</p>	Representation of number values	<ul style="list-style-type: none"> • Inconsistency in the representation of number values**
<p>Relevancy (Contextual): Is the data useful for the specified task? What kind of information is provided by a source? Does this information match the users' or system's requirements?</p>	Irrelevant information extracted	<ul style="list-style-type: none"> • Extraction of attributes containing layout information** • Redundant attribute values • Image related information* • Other irrelevant information

2) 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:

<http://aldda.b1.finki.ukim.mk/sparql>, <http://aldda.b1.finki.ukim.mk>. We tested the solution and deployment of the adopted tools (LODRefine, OpenLink Virtuoso, PoolParty Unified Views for a client from Libya. The PoolParty Unified Views (relevant for the speed layer in the Big Data architecture presented in Fig. 36) is considered for automation of the Extract-Transform-Load processes. The Unified Views' pipeline shall integrate also the custom quality assurance services discussed above.

V. Visualization and querying

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 [350]. 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 a diagram, formatted in different ways as shown in Figure 32. Custom visualization and query applications enable the user to interact with the data. To visualize the statistics about drug types and/or manufacturers, we used the exploratory spatial-temporal analysis (ESTA-LD) tool¹⁶⁵ [350]. The tool enables us to select the endpoint from where the data should be retrieved.

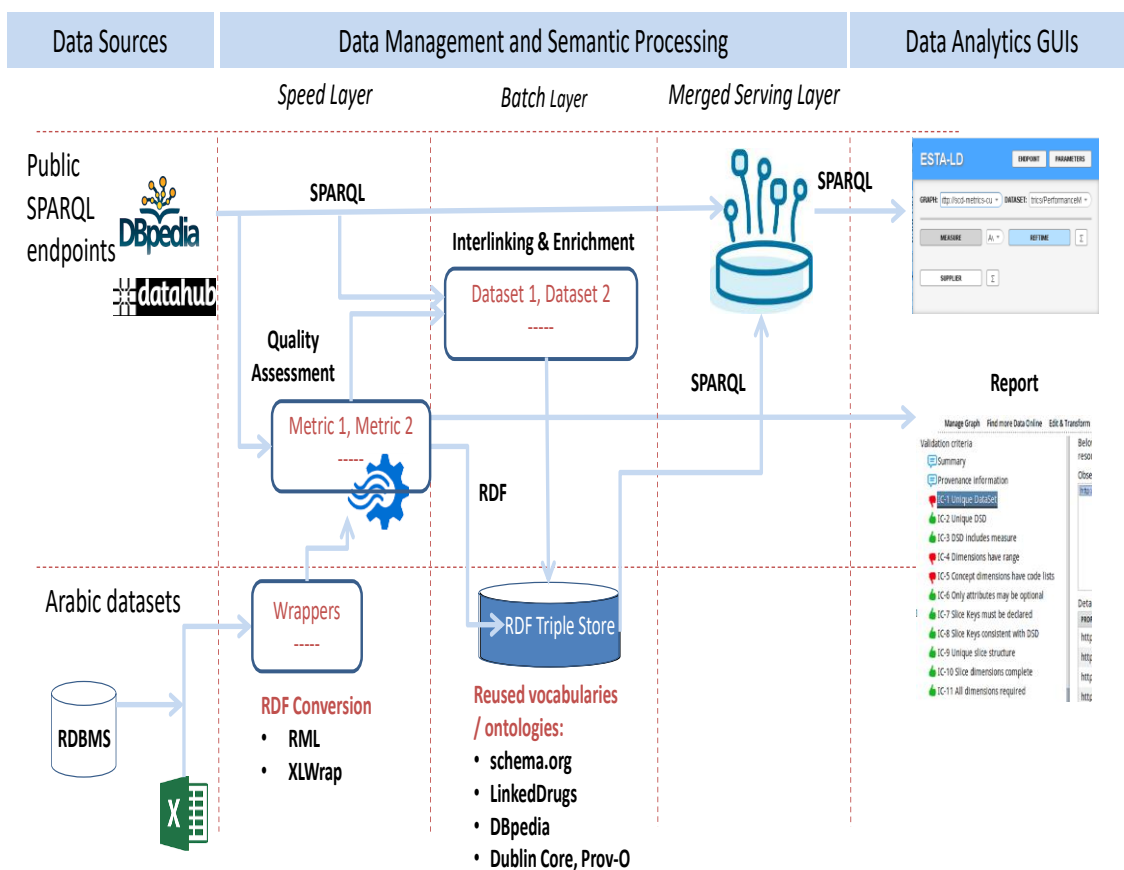


Figure 32: Knowledge graph visualization and querying

¹⁶⁵ <http://geoknow.imp.bg.ac.rs/ESTA-LD>

4.8 Summary

Most of the available Arabic drug datasets nowadays are still provided in a 2-star format and prepared in the English language, since the English language is widespread among physicians and pharmacists and also a predominant language in communications between physicians and pharmacists. To showcase the possibilities for large-scale integration of drug data, the candidate proposed a piloting methodology and tested the approach with datasets from Arabic countries. In the transformation process, the 2-star drug data was translated into a 5-star Linked Open Data format and interlinked 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.

The transformation process has been published in the journal paper

- *Guma Lakshen, Valentina Janev, and Sanja Vraneš. 2021. "Arabic Linked Drug Dataset: Consolidating and Publishing". Computer Science and Information Systems.2021. ComSIS Consortium. Volume 18, Issue 3, Pages: 729-748. <https://doi.org/10.2298/CSIS123456789X> and presented at the conference*
- *Guma Lakshen, Valentina Janev, and Sanja Vraneš. "Linking Open Drug Data: The Arabic dataset". 2019. "The Arabic dataset. In: Konjović, Z., Zdravković, M., Trajanović, M. (Eds.) ICIST 2019 Proceedings, pp.22-26.*

Chapter Five

Results and Analysis

CHAPTER FIVE – RESULTS AND ANALYSIS

The main reason for using Linked Open Data is to enable leveraging the value and usability of the dispersed data, in several use cases. As we have created the local ALDDA drug data interlinked and published with data from the LODD cloud, we can start querying our local data and continue crawling through the links to information published anywhere on the Web, thus extending the utilization possibilities of the data, and allowing the development of new types of applications over the data, serving our initial goal, that is, gaining value through linked open data.

In this chapter we will summarize the development of ALDDA-QA application and the process of consolidating and establishment of a Semantic Data Lake through the following sections:

Section 5.1 discusses the ALDDA-QA quality assessment process for ALDDA.

Section 5.2 presents business analytics on top of consolidated Arabic datasets. Via SPARQL endpoint queries we showcase the benefits from interlinking basic distributed drug datasets from different countries with mature well-known datasets (to enrich the original datasets for the sake of gaining additional value).

Section 5.3 points to issues with the Arabic DBpedia.

5.1 The Arabic Linked Drug Data Application Quality Assessment Architecture

The Quality Assessment component of the Arabic Linked Drug Data Application (ALDDA-QA) is a Java Web application based on the following frameworks:

The framework for testing the quality of DBpedia is a Java Web application based on the following frameworks:

- **Vaadin**, <https://vaadin.com/framework>, a Java framework for building Web applications used to implement the GUI (graphical user interface), while Sesame is used to execute SPARQL queries on the specified endpoint. The Web application can be deployed on any servlet container.
- **Sesame**, <https://sourceforge.net/projects/sesame/>: an open-source framework for querying and analyzing RDF data. It is an extensible architecture for efficient storage and expressive querying of large quantities of meta-data in RDF and RDF Schema. Sesame can be based on arbitrary repositories, ranging from traditional Data Base Management Systems to dedicated RDF triple stores. Sesame also implements a query engine for RQL, the most powerful RDF/RDF Schema query language to date. The

primary objective is to enable ALDDA-QA to operate on top of any SPARQL endpoint (a service for querying Linked Data) and enable the end-user to select the default endpoint and graph of interest. Further requests are to develop the ALDDA-QA as a standalone tool but, at the same time, be easily integrated into other similar environments, e.g., using the ESTA-LD tool for visualization of the statistics [350] (see Figure 33).

We implemented a stable and open-source version of the ALDDA-QA, which is a Java Web application based on AngularJS¹⁶⁶, an open-source Web applications framework designed to ease the development of single-page applications quality assessment framework that will allow the end-user to fully explore and, if possible, to repair the errors observed in the Arabic Linked Drug dataset.

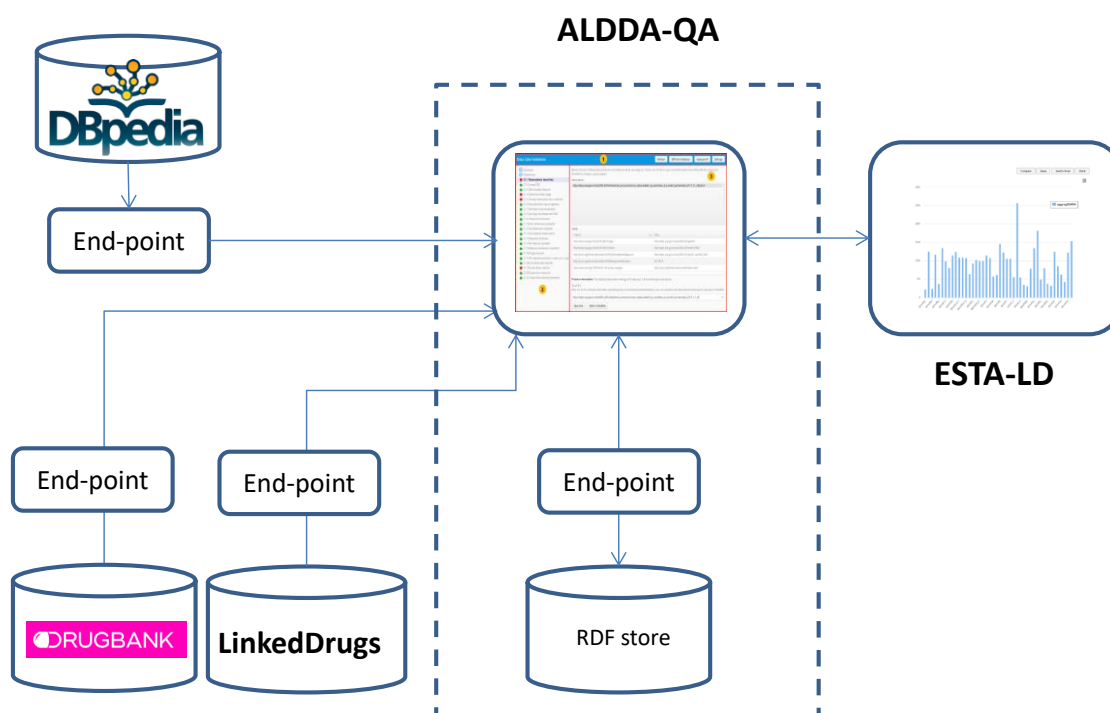


Figure 33: Quality Assessment Framework - Simplified illustration.

Data quality assessment, in this case, is a process whereby the initiator tests the data against a set of quality indicators, which in turn results in a fixed value that can be used to check whether the data is fit for use in the foreseen application. The ALDDA-QA is a Java Web application based on AngularJS¹⁶⁷, an open-source Web application framework designed to ease the development of single-page applications. Angular enables to separate data, views, and logic, providing dependency injection, while dynamic content is presented through two-way

¹⁶⁶<https://angularjs.org/>

¹⁶⁷<https://angularjs.org/>

data-binding that allows for the automatic synchronization of models and views. To animate content in the user interface, the Angular module ngAnimate¹⁶⁸ was used in tandem with jQuery¹⁶⁹. ngAnimate was used to introduce hooks that execute upon adding or removing elements, while jQuery was used for the actual animation.

5.2 Consolidating Arabic Open Drug Data

There exist a few Websites dealing with drugs such as WebTeb¹⁷⁰, altibbi¹⁷¹, and dwaprice¹⁷², etc, that give information about drugs such as brand names, usage, contraindications, prices, etc; but their data is not open and some information and only given based on registration and subscription (see chapter four for more details).

We presented SPARQL endpoint queries to visualize the benefits of interlinking basic distributed drug datasets from different countries with mature well-known datasets in the field to enrich the original datasets for the sake of gaining additional value. Some of these queries give answers to the questions in Section 4.1 above.

1. To know the number of distinct drugs in our ALDDA drug file and the number of interlinked drugs, we constructed the following two SPARQL endpoint queries.

Count all distinct drugs	Count all interlinked drugs
<pre> 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> } </pre>	<pre> 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 } </pre>
Output: 31906 distinct drugs	Output: 23971 interlinked drugs

Figure 34: Two SPARQL queries indicating the interlinkable drug's data

The output of the interlinking result shows that more than 75% (23971 out of 31906) of the drugs are interlinked with DBpedia and can obtain additional information in particular the Abstract information in the Arabic language that is required for the non-English speakers (intended users). This initial result encourages and serves the user to include abstract information from DBpedia not available in the ALDDA dataset, as it was prepared in the English language.

¹⁶⁸<https://docs.angularjs.org/api/ngAnimate>

¹⁶⁹<https://jquery.com/>

¹⁷⁰<https://www.Webteb.com/drug>

¹⁷¹<https://www.altibbi.com/الادوية>

¹⁷²<https://www.dwaprice.com/>

2. **Query1.** This query extracts abstract info from DBpedia in the 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 (partial):** The partial output extracts the Arabic language abstract information for the 'paclitaxel' drug from the DBpedia

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

3. **Query2:** Example question 'Fentanyl'

```
prefix dbo: <http://dbpedia.org/ontology/>
prefix drugbank: <http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugbank/>
prefix dbp: <http://dbpedia.org/ontology/>
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 .
?seeAlso dbo:wikiPageRevisionID ?wikiPageRevisionID .
OPTIONAL { ?seeAlso dbp:atcPrefix ?atcPrefix .}
OPTIONAL { ?seeAlso dbp:atcSuffix ?atcSuffix}
OPTIONAL { ?seeAlso owl:sameAs ?sameAs}
OPTIONAL { ?seeAlso dbp:synonyms ?synonyms}}
FILTER (?genericName = 'Fentanyl')
FILTER (langMatches(lang(?sameAs), "ar")) }
```

Output (partial): The partial output extracts the Arabic language abstract information for the 'Fentanyl' drug from the DBpedia

"الفينتانيل (بالإنجليزية) (Fentanyl): المعروف أيضا باسم (fentanil) والأسماء التجارية Sublimaze ، Actiq ، Durogesic ، Duragesic ، Fentora ، Onsolis ، Instanyl ، Abstral، وغيرها) هو من مسكنات المخدرات الاصطناعية الفعالة مع بداية سريعة ومدة قصيرة من العمل. وهو ناهض قوي على مستقبلات - μ الأفيونية. وتاريخيا، قد تم استخدامه لعلاج الألم المزمن ويستخدم عادة في مرحلة ما قبل الإجراءات الجراحية بمثابة مسكن للآلام وكمخدر في توليفة مع البنزوديازيبين. يعتبر الفينتانيل أقوى بـ 80 إلى 100 مرة من المورفين و بشكل تقريبي هو أقوى بـ 40 إلى 50 مرة من الهيروين المستخدم بشكل طبي (النقي 100%) صنع فينتانيل أول مرة من قبل باول جانسين في عام 1960. بعد الاكتشاف الطبي للبيثيديين في السنوات السابقة. طورت جانسين الفينتانيل عن طريق معايرة نظائر للدواء بيثيديين ذي البنية الكيميائية القريبة للفينتانيل بحثا عن الفاعلية الأفيونية. الاستخدام الواسع للفينتانيل أدى إلى إنتاج الفينتانيل سترات.

4. Query3: 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: The result of the query indicates that drug number 35704 'glimepiride' and the drug number 36482 'metformin and sulfonamides' have equivalent active ingredients and different generic names in two different countries.

	Drug1	Drug2
Drug Number	aldda.b1.finki.ukim.mk/lod/data/drugs#35704	aldda.b1.finki.ukim.mk/lod/data/drugs#36482
Generic Name	glimepiride	metformin and sulfonamides
Manufacturer Legal Name	Sadco	Benta Trading Co s.a.l.
Active Ingredient	Glimepiride	Glimepiride
CostFull	12415.0 L.L	31.04 SR
Address Country	LB	KSA

5. Query4: Drugs with different brand name comparisons.

```

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 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: In this query, the two drugs have the same generic name but with different brand names manufactured by two different manufacturers and with different dosage forms and prices in the same country.

	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

6. Query5: Display the drug description from DBpedia and Drug indication from Drugbank.

```
select distinct ?name str (?indication)str(?dbdesc)
where {
graph <http://aldda.b1.finki.ukim.mk/> {
aldda.b1.finki.ukim.mk/lod/data/drugs#35704
?pharmacy aldda.b1:pharmacyID '35704';
aldda:hasAvaliableMedicine ?drug.
?drug owl:sameAs ?alddaDrug.
}
graph <http://aldda.b1.finki.ukim.mk/> {
?alddaDrug rdfs:seeAlso ?dbdrug ;
drugbank:genericName ?name.
}
service <http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugbank/sparql>
{
?dbdrug drugbank:description ?dbdesc ;
owl:sameAs ?dpdrug.
?dbdrug drugbank:indication ?dbindication ;
owl:sameAs ?dpdrug.
}
service <http://dbpedia.org/sparql> {
?dpdrug dbpedia-owl:abstract ?dpdesc.
filter langMatches( lang(?dpdesc), "ar" )
}
}
```

Output: In this query, it is useful sometimes to obtain information regarding a particular drug, in this example ‘Glimepiride’ and display relevant information from drugbank and DBpedia in this example drug indication from the drugbank dataset and the description in Arabic from DBpedia. Additional attributes can be obtained through similar queries to enhance the value of the original dataset.

Name of Drug	Indication from Drug Bank	Description from DBpedia in Arabic (partial)
glimepiride	Glimepiride is indicated for the management of type 2 diabetes in adults as an adjunct to diet and exercise to improve glycemic control as monotherapy. It may also be indicated for use in combination with metformin or insulin to lower blood glucose in patients with type 2 diabetes whose high blood sugar levels cannot be controlled by diet and exercise in conjunction with an oral hypoglycemic (a drug used to lower blood sugar levels) agent alone	يخفض نسبة السكر في الدم عن طريق تحفيز إفراز الانسولين من البنكرياس، وهذا التأثير يعتمد على أداء خلايا بيتا في جُزيرات البنكرياس . الآلية التي يخفض فيها توليوتاميد السكر على المدى الطويل غير مفهومة . الإِستخدام المزمَن في مرضى السكري من النوع الثاني، وتخفيض تأثيرالسكر في الدم، يستمر على الرغم من الانخفاض التدريجي في استجابة الأنسولين المفرزة للدواء.

Similar queries can be devised to answer the remaining questions in section 2.3.

Figure 38 presents the ALDDA dashboard where users can select a drug from the list of drugs in the Combined Arabic dataset, select the required attributes from (see Table 17), select DrugBank attributes such as Summary, Background, Indication, Contraindications & Blackbox Warnings, Pharmacodynamics, Metabolism, Adverse Effects, Toxicity, and Food Interactions, then select whether or not include Arabic abstract for the drug from DBpedia.

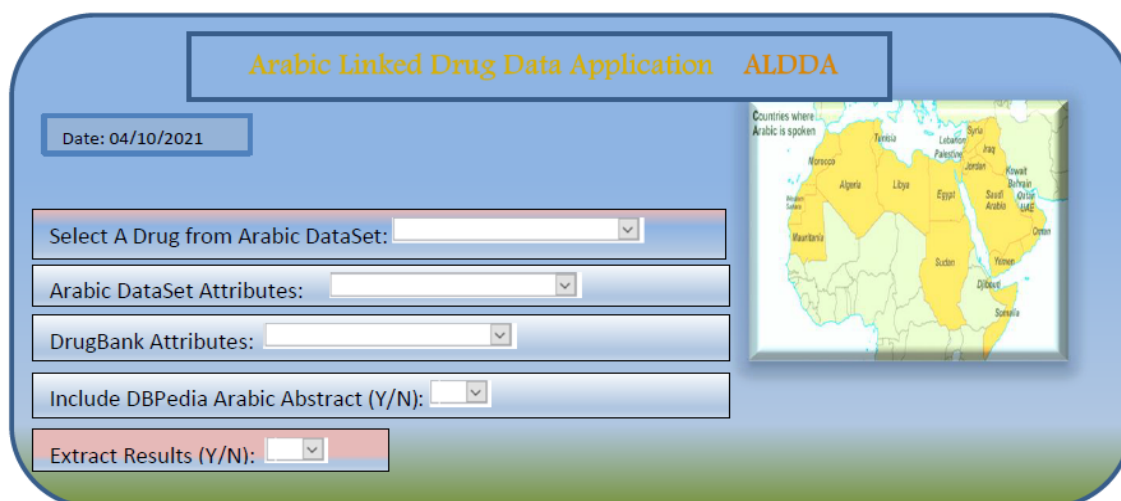


Figure 38: Screenshot of the Arabic Linked Drug Data Application ALDDA

Partial results of the search for the drug ‘Glimepiride’ would be:

Output:

Brand Name: GLIMEPIRIDE
Trade Name: AMARYL
Dosage Form: Tablet
Strength value: 1 mg , 2 mg, 3 mg

Manufacturer: sanofi-aventis- ITALY
Country: KSA, IRQ, LB, SYR

Background: First introduced in 1995, glimepiride is a member of the second-generation sulfonylurea (SU) drug class used for the management of type 2 diabetes mellitus (T2DM) to improve glycemic control. Type 2 diabetes is a metabolic disorder with increasing prevalences worldwide; it is characterized by insulin resistance in accordance with progressive β cell failure and long-term microvascular and macrovascular complications that lead to co-morbidities and mortalities.

Metabolism: Glimepiride is reported to undergo hepatic metabolism. Following either an intravenous or oral dose, glimepiride undergoes oxidative biotransformation mediated by CYP2C9 enzyme to form a major metabolite, cyclohexyl hydroxymethyl derivative (M1), that is pharmacologically active.

Toxicity: The oral LD50 value in rats is > 10000 mg/kg The intraperitoneal LD50 value in rats is reported to be 3950 mg/kg. Although glimepiride is reported to have fewer risks of hypoglycemia compared to other sulfonylureas such as glyburide, overdosage of glimepiride may result in severe hypoglycemia with coma, seizure, or other neurological impairment may occur. This can be treated with glucagon or intravenous glucose. Continued observation and additional carbohydrate intake may be necessary since hypoglycemia may recur after apparent clinical recovery

Food Interactions: Avoid alcohol. Acute and chronic alcohol intake may unpredictably affect the glucose-lowering action of glimepiride. Take with food. The manufacturer recommends administration with the first meal of the day.

يحتوى الدواء على المادة الفعالة جليمبيريد Glimepiride
جليمبيريد : هو دواء ينتمي إلى فئة من الأدوية تسمى سلفونيل يوريا Sulfonyl Ureas والتي تعمل على خفض سكر الدم عن طريق زيادة إفراز الأنسولين من البنكرياس.
البنكرياس يقوم بإنتاج الأنسولين والأنسولين هو مادة كيميائية يصنعها الجسم لنقل السكر (الجلوكوز) من الدم إلى داخل الخلايا وبمجرد دخول السكر إلى الخلايا ، يتم استخدامه كمصدر للطاقة. يحدث مرض السكري من النوع الثانى بسبب أن الجسم لا يُنتج ما يكفي من الأنسولين ، أو أن الخلايا لا تستجيب لتأثير الأنسولين بشكل صحيح وهذا ينتج عنه بقاء السكر في الدم وعدم دخوله لخلايا الجسم فيحدث ارتفاع فى نسبة السكر في الدم

5.3 Quality Assessment of Arabic DBpedia

Interlinking the Arabic Linked Drug dataset with DBpedia will further enhance the search capabilities in envisioned ALDDA applications. Therefore, in separate research (Paper 4 in the List of appended papers section), we investigated the quality of the Arabic DBpedia before consolidating it with DBpedia. DBpedia is heavily interlinked with other datasets and plays a central role in the linked open data cloud.

It is, therefore, a suitable data source for integration in cross-domain Linked Data applications, such as document annotation, faceted search, location-based information services, information extraction, and natural language processing services. However, incorrect metadata and incorrect or outdated data is a common problem when working with Linked Data in real-world applications [348]. Therefore, our goal related to DBpedia is to design a component that will:

- *Identify and explore errors e.g., incomplete or incorrectly extracted values from Wikipedia;*
- *Find irrelevant extraction of information and broken links;*
- *Find incorrectly extracted Datatypes.*
- *Identify representation problems and others.*

When it comes to the quality assessment of the DBpedia Arabic Chapter, there are problems specific to the Arabic language that result in:

- *Presentation of characters as symbols via Web browsers due to errors during the extraction process.*
- *Wrong values in numerical data, due to the use of Hindu numerals in some Arabic sources.*
- *Occurrence of different names for the same attribute, for instance, the birthdate attribute appears in various infoboxes by different names: one time as "(eng. birth date) الميلاذ تاريخ" another time as "(eng. delivery date) الولادة تاريخ", the third time as "(eng. birth) الميلاذ".*
- *The inconsistency of names between the infobox and its template; for instance, there is a template called "(eng. city) مدينة" while the infobox name is called "eng. city information مدينة معلومات"*
- *Geo-names templates formatting problems when placed in the infobox.*
- *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, we like to point to:

- *Wrong Wikipedia Infobox information; for example, the height of the 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.*
- *Mapping problems from Wikipedia, such as unavailability of infoboxes for many Arabic articles; for example, "Man-made River in Libya الصناعي النهر", which is considered as the biggest water pipeline project in the world, or not contain all the desired information.*
- *Object values are incompletely or incorrectly extracted.*
- *Data type incorrectly extracted.*
- *Some templates may be more abstract, thus cannot map to a specific class.*
- *Some templates are not used or missing inside the articles.*

5.4 Summary

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 the English language. The research showcases the benefits from the Linked Data approach, in particular the possibility of enriching the private datasets with selected open data such as DBpedia.

Based on the problems with the Arabic DBpedia, the candidate proposed a solution for the design of a quality assessment tool for Arabic-linked datasets. The quality assessment method is driven by the three dimensions that have been identified as relevant to the Arabic DBpedia, or Linked Data in general. The tests conducted in the research showed that the Arabic DBpedia dataset lacks continuous improvement, and it needs effective management to be used to efficiently enhance Arabic datasets. The results of the quality assessment of the Arabic DBpedia has been published as

- *Guma Abdulkhader Lakshen, Valentina Janev, and Sanja Vraneš. "Challenges in Quality Assessment of Arabic DBpedia". 2018. WIMS '18: Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics. June 2018. Article-No.: 15. Pages 1–4. <https://doi.org/10.1145/3227609.3227675>*

The main conclusion is that the Linked Data approach (1) contributes to the consolidation of the open datasets and 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.

Chapter Six

Conclusion and Future Directions

CHAPTER SIX - CONCLUSIONS AND FUTURE DIRECTIONS

In this chapter, we will review our research question and final goals to verify whether or not the goals were accomplished and whether the research questions were answered properly. Currently, to the best of our knowledge, comprehensive analysis and research of quality standards and quality assessment methods for Big and Linked Data are lacking in the available literature [197]. Enterprises that use the Linked Data approach are usually confronted with many challenges such as *heterogeneity and incompleteness, accuracy, relevancy, diversity of data sources, non-existing and approved data quality standards, lack of tools for error-handling, provenance management, and repairing of broken links.*

6.1 Analysis of the Linked Data Lifecycle

Public attention and literature awareness of the terms “*big data*”, “*open data*”, “*linked data*”, “*linked open data*”, “*public data*”, and “*government data*” etc., has grown tremendously in the last decades, for example, a generic search on a Google Search of the mentioned terms resulted in more than 245 million results. Our survey and analysis have shown that the Linked Data approach offers novel methods of publishing and binding data from various distributed sources and proposes a new spectrum of use case scenarios for developing creative implementations and services. It was observed clearly that, the drugs and the pharmaceutical domain are indeed embracing the Semantic Web technologies and the Linked Data principles, enabling critical information retrieval in the drugs domain, which is confronted with isolated data stores.

In this thesis, it was decided to use the RDF as the dataset format, because it is recommended by W3C, and has advantages, such as the provision of an extensible schema, self-describing data, de-referenceable URIs, and, as RDF links are typed, enables interoperability, structured, and safe linking of different datasets. Before starting converting the collected XLS to RDF format, target ontology was selected to describe the drugs contained in the drug availability dataset. We decided to use the Linked Drugs ontology, Schema.org¹⁷³ vocabulary, and DBpedia as they cover the needed properties and provide easier interlinking possibilities for further transformation.

The Web Ontology Language allows complex logical reasoning and consistency checking of RDF/OWL resources. These reasoning capabilities helped us to harmonize the heterogeneous data structures found in the input datasets. We transformed the selected drug data into five-star LOD and established relations in the RDF graph towards outside entities, including the DBpedia and DrugBank. The ‘owl:sameAs’ relation was selected to relate the drugs in the Arabic dataset with the entities in the Linked Drugs dataset and assumed that the two-drug descriptions refer to the same real-world entity. Most of the Web drug data in some Arabic

¹⁷³www.schema.org

countries are available as public 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 our proposal described above, we transformed the selected drug data into five-star linked open data and established relations in the RDF knowledge graph (31,906 drugs, 23,971 interlinked drugs to DBpedia and DrugBank, and 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 storing the knowledge graph, OpenLink virtuoso server (an Application Server Platform, version 06.01.3127), <https://github.com/openlink/virtuoso-opensource>. on Linux (x86_64-pc-Linux-gnu) was used.

For research purposes, the knowledge graph has been published via the SPARQL endpoint available at <http://aldda.b1.finki.ukim.mk/sparql>. We also provided use-cases that give examples of how the data from the Health Insurance Fund and DrugBank can be used, to provide application developers with mechanisms and ideas for retrieving distributed data in various formats.

6.2 Quality Analysis of Integrated Open Data

In this thesis, a comprehensive review was conducted on the innovative topic of big data and quality, which has gained a lot of attention and interest recently. There are lots of specific quality issues in the Linked Data lifecycle. In this thesis, the focus was both on data and metadata level, as well as functionalities for transformation and processing before visualization of consolidated data sets.

We can conclude that many technical challenges must be addressed first before the potential of interlinked data is realized fully in the business context of a pharmaceutical organization. The challenges include not just the obvious issues of scale, but also heterogeneity, lack of structure, error-handling, privacy, timeliness, and provenance, at all stages of the analysis pipeline from data acquisition to result in interpretation.

Based on the analysis of quality issues with DBpedia and the problems identified, we conclude that the most important dimensions to be taken into consideration for the consolidated Arabic Linked Drug Dataset are the following: *Accuracy*: triple incorrectly extracted, data type problems, errors in the implicit relationship between attributes, *Consistency*: representation of numerical values, and *Relevancy*: irrelevant information extracted. Different metrics were further defined, and Web services were implemented (see Table 19) to be used for data curation.

Hence, these challenges will require extensive testing with additional datasets to demonstrate the replicability of the developed prototype solution.

Table 19: Big Data challenges of and implemented functionalities related to quality

Challenge	Open Drug Data	Quality assessment functionalities
S1: Data volume is tremendous, the proportion of unstructured data in big data is very high and it is difficult to judge data quality within a reasonable amount of time.	The Arabic open datasets are small and it is possible to implement specific quality assessment services.	<ul style="list-style-type: none"> • accuracy and consistency checking • provenance and trustfulness of sources
S2: Data change very fast and the “timeliness” of data is very short	The Arabic open datasets have low velocity , hence no need for high-precision algorithms (or rules) to support the decision-making process.	
S3: The diversity of data sources brings abundant data types and complex data structures and increases the difficulty of data integration.	Variety was the focus of this thesis.	<ul style="list-style-type: none"> • validity checking to ensure conformity to agreed exchange standards • consistency checking to achieve a single version of the truth

6.3 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 (2012)* released as part of the Linked Data Integration Framework (LDIF; <http://sieve.wbsg.de/>) [313], *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 the RDFUnit Testing Suite (<https://github.com/AKSW/RDFUnit>) *Kontokostas et al.* enabled users 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>) [218], 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 20. These tools have not been tested with the Arabic DBpedia yet, an operation needed in our case study.

Table 20: Comparison of open-source quality assessment tools according to several attributes

Tool	Extensibility	Last Update	Collaboration	Cleaning Support
RDFUnit	SPARQL	03/2018	×	×
Luzzu	JAVA, LQML	07/2017	×	×
TripleCheckMate	×	03/2017	√	×
Laundromat	SPARQL	05/2018	√	√
Sieve	XML	2014	×	√

In conclusion, we can state that modern organizations can derive significant added value from embracing knowledge management principles to promote a smooth flow, sharing, and re-using of both internal and external knowledge and information. However, since R&D organizations' innovation charter demands a focus different from that of other types of organizations, specifically, to nurture open access to human resources' extensive knowledge and experience, both explicit and tacit, significant adjustment of standard knowledge management solutions and practices are necessary to suit their needs. These adjustments have been described in this Ph.D. thesis.

Most of the available drug datasets in the Arabic countries nowadays are still provided in a 2-star format in the English language since the English language is widespread among physicians and pharmacists and also a predominant language in communications between physicians and pharmacists. To showcase the possibilities for large-scale integration of drug data, we proposed a piloting methodology and tested the approach with datasets from Arabic countries.

We presented the transformation process of 2-star drug data into a 5-star Linked Open Data with DrugBank and DBpedia. The thesis showcases benefits from the Linked Data approach and for the first time discusses the issues with drug data from Arabic countries (author selected four-drug data files from four different Arabic countries, Iraq, Syria, Saudi Arabia, and Lebanon).

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

implementation of a stable and open-source version of a Java Web application that will allow the end-user to fully explore and assess the quality of the consolidated dataset, and if possible, to repair the errors observed in the Arabic Linked Drug dataset.

The thesis showcases the benefits from the Linked Data approach, in particular the possibility of enriching the private datasets with selected open data such as DBpedia and Drugbank. The main conclusion is that the Linked Data approach: i) contributes to the standardization on the metadata level and the Semantic interoperability; ii) 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. We strongly believe that quality issues in the drug industry in the Arab countries still need further study and evaluation.

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 their deficiencies in the English language. The main contributions can be summarized as follows:

- *This work introduced a modified process model based on previous methodologies.*
- *It is recommended to use quality assessment services in the process of selecting open data, its transformation, and processing to ensure that the process is conducted in a high-quality manner.*
- *For the first time, the issues with drug data from Arabic countries were discussed based on the selected four-drug data files from four different Arabic countries, Iraq, Syria, Saudi Arabia, and Lebanon.*

The described novel methodologies and applications are fully transferable to future data sets which might become available in the Arabic language.

BIBLIOGRAPHY

- [1] D. AnHai, H. Alon, and I. Zachary, *Principles of Data Integration*. 2012.
- [2] A. Haug, F. Zachariassen, and D. van Liempd, "The costs of poor data quality," *J. Ind. Eng. Manag.*, vol. 4, no. 2, pp. 168–193, 2011.
- [3] N. Laranjeiro, S. N. Soydemir, and J. Bernardino, "A Survey on Data Quality: Classifying Poor Data," in *Proceedings - 2015 IEEE 21st Pacific Rim International Symposium on Dependable Computing, PRDC 2015*, 2016.
- [4] D. Loshin, "Evaluating the Business Impacts of Poor Data Quality," *Softw. Eng. Inst.*, no. 301, pp. 1–10, 2010.
- [5] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, and S. Auer, "Quality assessment for Linked Data: A Survey," in *Semantic Web*, 2016.
- [6] A. Zaveri, A. Maurino, and L. B. Equille, "Web data quality: Current state and new challenges," *Int. J. Semant. Web Inf. Syst.*, vol. 10, no. 2, pp. 1–6, 2014.
- [7] KPMG International, "2016 Global CEO Outlook," pp. 1–8, 2016.
- [8] D. Adams, H. Behnke, G. Bonnette, J. Francica, and M. Goetz, "The Data Differentiator," 2017.
- [9] C. Cichy and S. Rass, "An overview of data quality frameworks," *IEEE Access*, vol. 7, pp. 24634–24648, 2019.
- [10] C. Bizer, T. Heath, and T. Berners-Lee, "Linked data - The story so far," *Int. J. Semant. Web Inf. Syst.*, 2009.
- [11] G. Antoniou and F. Van Harmelen, *A Semantic Web Primer*. .
- [12] T. Berners-Lee, J. Hendler, and O. Lassila, "The semantic web," *Scientific American*. 2001.
- [13] N. Shadbolt, W. Hall, and T. Berners-Lee, "The semantic web revisited," *IEEE Intelligent Systems*. 2006.
- [14] V. R. Benjamins, J. Contreras, O. Corcho, and A. Gomez-Perez, "Six challenges for the semantic web," in *First International Semantic Web Conference, ISWC2002*, 2002.
- [15] L. Sebastian-Coleman, "Measuring Data Quality for Ongoing Improvement," *Meas. Data Qual. Ongoing Improv.*, 2013.
- [16] M. Ge, "Information quality assessment and effects on inventory decision-making," *Dublin City Univ.*, no. September, pp. 1–185, 2009.
- [17] N. Elgendy and A. Elragal, "Big data analytics: A literature review paper," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2014.
- [18] A. R. Syed, K. Gillela, and C. Venugopal, "The Future Revolution on Big Data," *Int. J. Adv. Res. Comput. Commun. Eng.*, 2013.
- [19] F. Tekiner and J. A. Keane, "Big data framework," in *Proceedings - 2013 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2013*, 2013.
- [20] P. Almeida and J. Bernardino, "A comprehensive overview of open source big data platforms and frameworks," *Serv. Trans. Big Data*, 2015.
- [21] M. D. Assunção, R. N. Calheiros, S. Bianchi, M. A. S. Netto, and R. Buyya, "Big Data computing and clouds: Trends and future directions," *J. Parallel Distrib. Comput.*, 2015.
- [22] T. Davies, "Open Data : Infrastructures and ecosystems," *Open Data Res.*, pp. 1–6, 2011.
- [23] Michel Foucault, *The Order of Things*. London and New York, 2013.
- [24] T. Berners-Lee, "Plenary talk by Tim BL at WWWF94: Overview." 1994.
- [25] I. S. Web, S. Bechhofer, P. De Madrid, and A. Gangemi, "Introduction to the Semantic Web Knowledge Representation - Ontologies What is Knowledge Representation," vol. 2006, no. 2, 2006.
- [26] J. Hendler, "Is there an intelligent agent in your future?," *Nature*, 1999.
- [27] P. V Patil and M. N. Nachappa, "Semantic Web and Web Mining For Ontology Building," vol. 6, no.

- 8, pp. 154–157, 2019.
- [28] L. Sauer mann, R. Cyganiak, D. Ayers, and M. Völkel, “Cool URIs for the Semantic Web. W3C Interest Group Note,” *W3C*, 2008.
- [29] G. Antoniou, E. Franconi, and F. Van Harmelen, “Introduction to Semantic Web Ontology Languages,” pp. 1–21, 2005.
- [30] C. Burlison, “Introduction to the Semantic Web Vision and Technologies - Part 1 - Overview - Blog - Semantic Focus - The Semantic Web, Semantic Web technology and computational semantics,” 2007.
- [31] J. A. H. John Domingue, Dieter Fensel, *Handbook of Semantic Web*. 2011.
- [32] O. Lassila and R. R. Swick, “Resource Description Framework (RDF) Model and Syntax Specification,” no. 19990222. 1999.
- [33] H. Jabeen, D. Graux, and G. Sejdiu, “Scalable knowledge graph processing using sansa,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2020.
- [34] X. Du, “Semantic service description framework for efficient service discovery and composition,” 2009.
- [35] K. Nath, S. Dhar, and S. Basishtha, “Web 1.0 to Web 3.0 - Evolution of the Web and its various challenges,” *ICROIT 2014 - Proc. 2014 Int. Conf. Reliab. Optim. Inf. Technol.*, pp. 86–89, 2014.
- [36] J. D. King, “Search Engine Content Analysis,” 2008.
- [37] S. Aghaei, M. A. Nematbakhsh, and H. K. Farsani, “E Volution of the W Orld W Ide W Eb : From,” *Int. J. Web Semant. Technol.*, vol. 3, no. 1, pp. 1–10, 2012.
- [38] K. D. Foote, “A Brief History of Big Data - DATAVERSITY.” 2017.
- [39] M. Fourment and M. R. Gillings, “A comparison of common programming languages used in bioinformatics,” *BMC Bioinformatics*, vol. 9, no. 6, pp. 8096–8100, 2008.
- [40] T. Berners-Lee, “The World Wide Web: A very short personal history,” *W3.Org*. 1998.
- [41] P. A. Bonatti, A. Hogan, A. Polleres, and L. Sauro, “Robust and scalable Linked Data reasoning incorporating provenance and trust annotations,” *J. Web Semant.*, 2011.
- [42] N. Bessis, E. Assimakopoulou, M. E. Aydin, and F. Xhafa, “Utilizing next generation emerging technologies for enabling collective computational intelligence in disaster management,” *Stud. Comput. Intell.*, vol. 352, pp. 503–526, 2011.
- [43] W3C OWL Working Group, “OWL 2 Web Ontology Language Document Overview,” *OWL 2 Web Ontol. Lang.*, no. December, pp. 1–7, 2012.
- [44] B. Glimm and C. Ogbuji, “SPARQL 1.1 Entailment Regimes W3C Recommendation 21 March 2013,” *W3C*. 2013.
- [45] S. Muñoz, J. Pérez, and C. Gutierrez, “Minimal deductive systems for RDF,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 4519 LNCS, pp. 53–67, 2007.
- [46] I. Kollia, B. Glimm, and I. Horrocks, “SPARQL query answering over OWL ontologies,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2011.
- [47] F. Manola and E. M. (editors), “{RDF Primer}.” 2004.
- [48] A. Dimou, “High Quality Linked Data Generation from Heterogeneous Data,” no. november, 2017.
- [49] J. Gao, “Linked Data Based Enterprise Data Integration,” 2011.
- [50] P. V Biron and A. Malhotra, “XML Schema Part 2: Datatypes Second Edition,” *W3C Recommendation*. 2004.
- [51] H. Alvestrand, “RFC 3066 - Tags for the Identification of Languages.” Cisco Systems, 2001.
- [52] D. Brickley and R. V. Guha, “RDF Vocabulary Description Language 1.0: RDF Schema, W3C Recommendation 10 February 2004,” *W3C*, 2004. .
- [53] S. R. M. Zeebaree, A. Al-Zebari, K. Jacksi, and A. Selamat, “Designing an ontology of E-learning system for duhok polytechnic university using protégé OWL tool,” *J. Adv. Res. Dyn. Control Syst.*, vol. 11, no. 5, pp. 24–37, 2019.
- [54] S. Muñoz, J. Pérez, and C. Gutierrez, “Simple and Efficient Minimal RDFS,” *J. Web Semant.*, 2009.

- [55] S. C. Shapiro, "Knowledge Representation: Logical, Philosophical, and Computational Foundations John F. Sowa Pacific Grove, CA: Brooks/Cole, 2000, xiv+594 pp; hardbound, ISBN 0-534-94965-7, \$67.95," *Comput. Linguist.*, 2001.
- [56] R. Studer and V. R. Benjamins, "Knowledge Engineering : Principles and Methods," vol. 25, no. February 2018, pp. 161–197, 1998.
- [57] M. P. S. Bhatia, A. Kumar, and R. Beniwal, "Ontologies for software engineering: Past, present and future," *Indian J. Sci. Technol.*, 2016.
- [58] A. T. Zouhair Rimale, EL Habib Benlahmar, "A Semantic Learning Object (SLO) Web-Editor based on Web Ontology Language (OWL) using a New OWL2XSLO Approach," vol. 7, no. 12, pp. 315–320, 2016.
- [59] P. Yongyuth *et al.*, "The AGROVOC concept server workbench: A collaborative tool for managing multilingual knowledge," *Retrieved on December*, vol. 10, p. 2010, 2008.
- [60] S. K. Raffat, M. Sarim, and S. Iqbal, "OBO edit: A tool for classifying the Biological Data.," *FUUAST J. Biol.*, vol. 5, no. 2, pp. 249–255, 2015.
- [61] A. Kalyanpur, B. Parsia, E. Sirin, B. C. Grau, and J. Hendler, "Swoop: A Web Ontology Editing Browser," *J. Web Semant.*, vol. 4, no. 2, pp. 144–153, Jun. 2006.
- [62] E. Lee, N. Harris, M. Gibson, R. Chetty, and S. Lewis, "Apollo: A community resource for genome annotation editing," *Bioinformatics*, vol. 25, no. 14, pp. 1836–1837, 2009.
- [63] E. S. Alatrish, "Comparison of Ontology Editors," *e-RAFJ. Comput.*, vol. 4, pp. 23–38, 2012.
- [64] A. S. Al-Wabil and H. S. Al-Khalifa, "The Arabic language and the semantic web : Challenges and opportunities," *CiteSeerX*, p. 10, 2007.
- [65] S. Tomic, A. Fensel, and T. Pellegrini, "SESAME demonstrator: Ontologies, services and policies for energy efficiency," *ACM Int. Conf. Proceeding Ser.*, pp. 1–4, 2010.
- [66] K. Balachandar, E. Thirumagal, D. Aishwarya, and R. Rajkumar, "Ontology Mapping Techniques and Approaches," *Int. J. Comput. Appl.*, vol. 65, no. 24, pp. 13–20, 2013.
- [67] T. Slimani, "Ontology Development: A Comparing Study on Tools, Languages and Formalisms," *Indian J. Sci. Technol.*, vol. 8, no. 24, 2015.
- [68] L. Ehrlinger and W. Wöß, "Towards a definition of knowledge graphs," *CEUR Workshop Proc.*, vol. 1695, no. September 2016, 2016.
- [69] A. Hogan *et al.*, "Knowledge graphs," *arXiv*. 2020.
- [70] M. Krötzsch, "Ontologies for knowledge graphs?," in *CEUR Workshop Proceedings*, 2017.
- [71] M. Calaresu and A. Shiri, "Understanding semantic web: a conceptual model: For Authors Understanding Semantic Web: A Conceptual Model," *Libr. Rev.*, vol. 64, no. 12, 2015.
- [72] H. Hedden, "Controlled vocabularies, thesauri, and taxonomies," *Index. Int. J. Index.*, 2008.
- [73] S. Arora and B. Boaz, *Computational Complexity: A Modern Approach*, no. January. 2007.
- [74] B. C. Grau, I. Horrocks, B. Motik, B. Parsia, P. Patel-Schneider, and U. Sattler, "OWL 2: The next step for OWL," *Web Semant.*, vol. 6, no. 4, pp. 309–322, 2008.
- [75] W. M. Spears, "Pushing the envelope," *Physicomimetics Physics-Based Swarm Intell.*, vol. 9783642228, pp. 93–125, 2012.
- [76] A. Artale, D. Calvanese, R. Kontchakov, and M. Zakharyashev, "The DL-Lite family and relations," *J. Artif. Intell. Res.*, vol. 36, pp. 1–69, 2009.
- [77] H. Zhu, S. Madnick, Y. Lee, and R. Wang, *Data and Information Quality Research*. 2014.
- [78] M. M. Aref and Z. Zhou, "The Ontology Web Language (OWL) for a multi-agent understating system," *2005 Int. Conf. Integr. Knowl. Intensive Multi-Agent Syst. KIMAS'05 Model. Explor. Eng.*, vol. 2005, pp. 586–591, 2005.
- [79] S. Mehla and S. Jain, *Rule languages for the semantic web*, vol. 755. Springer Singapore, 2019.
- [80] C. Bizer and A. Schultz, "The R2R framework: Publishing and discovering mappings on the web," *CEUR Workshop Proc.*, vol. 665, 2010.
- [81] N. Gur, L. Díaz Sanchez, and T. Kauppinen, "GI Systems for Public Health with an Ontology Based Approach," *Bridg. Geogr. Inf. Sci. Int. Agil. Conf. Avignon Fr. April 24- 27, 2012*, pp. 86–91, 2012.
- [82] A. Zaveri, J. Lehmann, S. Auer, M. M. Hassan, M. A. Sherif, and M. Martin, "Publishing and interlinking the global health observatory dataset," *Semant. Web*, 2013.

- [83] Doug Laney, "3D Data Management: Controlling Data Volume, Velocity, and Variety," *META Gr.*, no. February 2001, 2001.
- [84] G. Shankaranarayan, M. Ziad, and R. Y. Wang, "Managing data quality in dynamic decision environments: An information product approach," *J. Database Manag.*, 2003.
- [85] D. Garlasu, V. Sandulescu, and M. Marinescu, "A Big Data implementation based on Grid Computing," *11th Roedunet Int. Conf. (RoEduNet), Sinaia*, pp. 1–4, 2013.
- [86] J. Manyika *et al.*, "Big data : The next frontier for innovation , competition , and productivity," 2011.
- [87] Gartner Inc., "What Is Big Data? - Gartner IT Glossary - Big Data," *Gartner IT Glossary*, 2013. .
- [88] M. A. Beyer and D. Laney, "The importance of 'big data': a definition," *Stamford, CT Gart.*, 2012.
- [89] Zikopoulos Paul and Eaton Chris, *Understanding Big Data*, vol. 1, no. 2019.
- [90] A. Patrizio, "Expect 175 zettabytes of data worldwide by 2025," *NetworkWorld*, 2019. .
- [91] Ying Lin, "10 Internet Statistics Every Marketer Should Know in 2020 [Infographic]," 2019. .
- [92] S. Kemp, "Digital 2020: Global Digital Overview — DataReportal – Global Digital Insights," *Datareportal.com*, 2020. [Online]. Available: <https://datareportal.com/reports/digital-2020-global-digital-overview%0Ahttps://datareportal.com/reports/digital-2020-global-digital-overview%0Ahttps://datareportal.com/reports/digital-2020-ecuador%0Ahttps://datareportal.com/reports/digital-2020-global-di>.
- [93] MIKE2.0, "Open Framework, Information Management Strategy & Collaborative Governance | Data & Social Methodology - MIKE2.0 Methodology." .
- [94] Wikipedia, "Big data - Wikipedia," *Wikipedia, The Free Encyclopedia*, 2019. [Online]. Available: http://en.wikipedia.org/wiki/Big_data.
- [95] NESSI, "Big Data_ A New World for Opportunities," *insideBIGDATA*, p. 25, 2012.
- [96] Microsoft, "The Big Bang: How the Big Data Explosion Is Changing the World," *Microsoft News Center*. pp. 1–35, 2013.
- [97] R. L. Villars, C. W. Olofson, and M. Eastwood, "Big Data: What It is and Why You Should Care," *IDC White Pap.*, pp. 7–8, 2011.
- [98] A. Brust, "Big Data: Defining its definition," *ZDNet*, 2012.
- [99] A. Jacobs, "The Pathologies of Big Data," 2009.
- [100] M. Chen, S. Mao, and Y. Liu, "Big data: A survey," *Mob. Networks Appl.*, vol. 19, no. 2, pp. 171–209, 2014.
- [101] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," *Int. J. Inf. Manage.*, vol. 35, no. 2, pp. 137–144, 2015.
- [102] C. W. Tsai, C. F. Lai, H. C. Chao, and A. V. Vasilakos, "Big data analytics: a survey," *J. Big Data*, vol. 2, no. 1, pp. 1–32, 2015.
- [103] B. Klievink, B. J. Romijn, S. Cunningham, and H. de Bruijn, "Big data in the public sector: Uncertainties and readiness," *Inf. Syst. Front.*, vol. 19, no. 2, pp. 267–283, 2017.
- [104] M. Loukides, "What is data science – O'Reilly." O'Reilly Radar, 2010.
- [105] M. Stonebraker, "What Does ' Big Data ' Mean and Who Will Win ?," *Xldb*, 2012.
- [106] A. De Mauroandrea, M. Greco, M. Grimaldim, and V. Table, "What is Big Data ? A Consensual Definition and a Review of Key Research Topics," vol. 97, 2015.
- [107] M. Lněnička and J. Komárková, "The Impact of Cloud Computing and Open (Big) Data on the Enterprise Architecture Framework," in *Proceedings of the 26th International Business-Information-Management-Association Conference (pp. 1679–1683)*. Norristown: IBIMA., 2015, no. June, pp. 1679–1683.
- [108] G. Vossen, "Big data as the new enabler in business and other intelligence," *Vietnam J. Comput. Sci.*, vol. 1, no. 1, pp. 3–14, 2014.
- [109] B. Kajruba, G. Lakshen, and J. Velickovic, "The impact of big data analysis in enhancing sustainable development goals," *1st Conf. Sustain. Dev. from Econ. Perspect.*, vol. 1, no. 1, pp. 1–19, 2021.
- [110] S. Suthaharan, "Big Data Classification: Problems and Challenges in Network Intrusion Prediction with Machine Learning," vol. 41 (4), pp. 70–73, 2014.
- [111] George Firican, "The 10 Vs of Big Data | Transforming Data with Intelligence," *TDWI*, 2017.

- [Online]. Available: <https://tdwi.org/articles/2017/02/08/10-vs-of-big-data.aspx>.
- [112] J. Mervis, "U.S. science policy: Agencies rally to tackle big data," *Science (80-.)*, vol. 336, no. 6077, pp. 22–22, Apr. 2012.
- [113] B. Brown, M. Chui, and J. Manyika, "Are you ready for the era of 'big data'? | McKinsey & Company," *McKinsey Q.*, 2011.
- [114] C. W. Choo, *The Knowing Organization: How Organizations Use Information to Construct Meaning, Create Knowledge, and Make Decisions*. 2007.
- [115] M. E. Porter, *competitive advantage: Creating and Sustaining Superior Peifonnance*. 1985.
- [116] E. Curry, "The Big Data Value Chain: Definitions, Concepts, and Theoretical Approaches," in *New Horizons for a Data-Driven Economy*, Cham: Springer International Publishing, 2016, pp. 29–37.
- [117] G. Lakshen, S. Vranes, and V. Janev, "Big data and quality: A literature review," in *24th Telecommunications Forum, TELFOR 2016*, 2017.
- [118] V. Marx, "The big challenges of big data," *Nature*, 2013.
- [119] A. Labrinidis and H. V. Jagadish, "Challenges and opportunities with big data," *Proc. VLDB Endow.*, 2012.
- [120] A. Katal, M. Wazid, and R. H. Goudar, "Big data: Issues, challenges, tools and Good practices," in *2013 6th International Conference on Contemporary Computing, IC3 2013*, 2013.
- [121] L. Cai and Y. Zhu, "The challenges of data quality and data quality assessment in the big data era," *Data Sci. J.*, vol. 14, pp. 1–10, 2015.
- [122] I. Lee, "Big data: Dimensions, evolution, impacts, and challenges," *Bus. Horiz.*, 2017.
- [123] P. Kaur and A. A. Monga, "Managing Big Data: A Step towards Huge Data Security," *Int. J. Wirel. Microw. Technol.*, vol. 6, no. 2, pp. 10–20, 2016.
- [124] Z. K. Lawal and R. Y. Zakari, "A review: Issues and Challenges in Big Data from Analytic and Storage perspectives," *Int. J. Eng. Comput. Sci.*, no. December, 2016.
- [125] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic, "Deep learning applications and challenges in big data analytics," *J. Big Data*, vol. 2, no. 1, pp. 1–21, 2015.
- [126] A. Wulff and C. Wunck, "Integration of business process management and Big Data technologies," *Proc. Int. Conf. Ind. Eng. Oper. Manag.*, vol. 8-10 March, p. 2316, 2016.
- [127] A. Rula, A. Maurino, and C. Batini, "Data Quality Issues in Linked Open Data," 2016.
- [128] D. P. Acharjya, "A Survey on Big Data Analytics : Challenges , Open Research Issues and Tools," vol. 7, no. 2, 2016.
- [129] D. Robb, "Top Ten Big Data Storage Tools." .
- [130] H. Palovská, "What Can NoSQL Serve an Enterprise," *J. Syst. Integr.*, pp. 44–49, 2015.
- [131] C. Batini and M. Scannapieco, *Data and Information Quality, Dimensions, Principles and Techniques*. 2016.
- [132] L. Cai and Y. Zhu, "The challenges of data quality and data quality assessment in the big data era," in *Data Science Journal*, 2015.
- [133] A. Nikiforova, *Definition and Evaluation of Data Quality: User-Oriented Data Object-Driven Approach to Data Quality Assessment*, vol. 8, no. 3. 2020.
- [134] M. Scannapieco and T. Catarci, "Data Quality under the Computer Science perspective," *Comput. Eng.*, vol. 2, no. 2, pp. 1–12, 2002.
- [135] R. Y. Wang and D. M. Strong, "Beyond Accuracy : What Data Quality Means to Data Consumers," *Manag. Inf. Syst.*, vol. 12, no. 4, pp. 5–33, 2013.
- [136] H. E. Miller, "The Multiple Dimensions of Information Quality," *Inf. Syst. Manag.*, vol. 13(2), no. March 1996, pp. 79–82, 2015.
- [137] M. Chen, M. Song, J. Han, and E. Haihong, "Survey on data quality," in *Proceedings of the 2012 World Congress on Information and Communication Technologies, WICT 2012*, 2012.
- [138] P. Zhang, F. Xiong, J. Gao, and J. Wang, "Data quality in big data processing: Issues, solutions and open problems," in *2017 IEEE SmartWorld Ubiquitous Intelligence and Computing, Advanced and Trusted Computed, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People and Smart City Innovation, SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI 2017 - , 2018*.

- [139] C. Batini, A. Rula, M. Scannapieco, and G. Viscusi, "From data quality to big data quality," *J. Database Manag.*, 2015.
- [140] E. Rahm and H. Do, "Data Cleaning : Problems and Current Approaches," pp. 1–11, 2000.
- [141] C. Bizer and R. Cyganiak, "Quality-driven information filtering using the WIQA policy framework," *Web Semant.*, 2009.
- [142] C. Cappiello, W. Samá, and M. Vitali, "Quality awareness for a successful big data exploitation," in *ACM International Conference Proceeding Series*, 2018.
- [143] M. Jovanovik and D. Trajanov, "Consolidating drug data on a global scale using Linked Data," *J. Biomed. Semantics*, vol. 8, no. 1, pp. 1–24, 2017.
- [144] A. C. N. Ngomo, S. Auer, J. Lehmann, and A. Zaveri, "Introduction to linked data and its lifecycle on the web," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, 2014.
- [145] J. El-khoury, D. Gürdür, and M. Nyberg, "A Model-Driven Engineering Approach to Software Tool Interoperability based on Linked Data," *Int. J. Adv. Softw.*, vol. 9, no. 3 & 4, pp. 248–259, 2016.
- [146] M. B. Josko and E. Ferreira, "Visualization properties for data quality visual assessment: An exploratory case study," 2016.
- [147] B. W. Boehm, "Software Engineering Economics," *IEEE Trans. Softw. Eng.*, 1984.
- [148] J. Debattista, C. Lange, S. Auer, and D. Cortis, "Evaluating the quality of the LOD cloud: An empirical investigation," *Semant. Web*, vol. 9, no. 6, pp. 859–901, 2018.
- [149] A. Hogan, A. Harth, A. Passant, S. Decker, and A. Polleres, "Weaving the pedantic Web," in *CEUR Workshop Proceedings*, 2010.
- [150] N. Abdullah, S. A. Ismail, S. Sophiayati, and S. M. Sam, "Data quality in big data: A review," *Int. J. Adv. Soft Comput. its Appl.*, vol. 7, no. Specialissue3, pp. 16–27, 2015.
- [151] K. Adam, I. Hammad, M. Adam, I. Fakharaldien, and M. A. Majid, "Big Data Analysis and Storage," pp. 648–659, 2015.
- [152] M. Padgavankar and S. Gupta, "Big data storage and challenges," *Int. J. Comput. Sci. Inf. Technol.*, vol. 5, no. 2, pp. 2218–2223, 2014.
- [153] C.-W. Tsai, C.-F. Lai, H.-C. Chao, and A. V. Vasilakos, "Big data analytics: a survey," *J. Big Data*, vol. 2, no. 21, 2015.
- [154] Ravindra Phule;Madhav Ingle, "A Survey on Scalable Big Data Analytics Platform," *i-manager's J. Cloud Comput.*, vol. 2, no. 4, pp. 43–48, 2015.
- [155] M. Färber, F. Bartscherer, C. Menne, and A. Rettinger, "Linked data quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO," *Semant. Web*, vol. 9, no. 1, pp. 77–129, 2018.
- [156] B. Walshe, R. Brennan, and D. O'Sullivan, "Bayes-ReCCE: A Bayesian model for detecting restriction class correspondences in linked open data knowledge bases," *Int. J. Semant. Web Inf. Syst.*, vol. 12, no. 2, pp. 25–52, 2016.
- [157] B. Adrian, "The 13 types Of data," *Forbes*. 2018.
- [158] M. Janssen, Y. Charalabidis, and A. Zuiderwijk, "Benefits, Adoption Barriers and Myths of Open Data and Open Government," *Inf. Syst. Manag.*, vol. 29, no. 4, pp. 258–268, 2012.
- [159] C. P. Geiger and J. Von Lucke, "Open Government and (Linked) (Open) (Government) (Data)," *JeDEM - eJournal eDemocracy Open Gov.*, vol. 4, no. 2, pp. 265–278, 2012.
- [160] D. S. Sayogo, T. A. Pardo, and M. Cook, "A framework for benchmarking open government data efforts," *Proc. Annu. Hawaii Int. Conf. Syst. Sci.*, no. May 2010, pp. 1896–1905, 2014.
- [161] J. Attard, F. Orlandi, S. Scerri, and S. Auer, "A systematic review of open government data initiatives," *Gov. Inf. Q.*, vol. 32, no. 4, pp. 399–418, 2015.
- [162] E. Ruijter, S. Grimmelikhuijsen, and A. Meijer, "Open data for democracy: Developing a theoretical framework for open data use," *Gov. Inf. Q.*, vol. 34, no. 1, pp. 45–52, 2017.
- [163] E. Barry and F. Bannister, "Barriers to open data release: A view from the top," *Inf. Polity*, vol. 19, no. 1–2, pp. 129–152, 2014.
- [164] K. Hardy and A. Maurushat, "Opening up government data for Big Data analysis and public benefit," *Comput. Law Secur. Rev.*, vol. 33, no. 1, pp. 30–37, 2017.
- [165] J. Kučera, D. Chlapek, and M. Nečaský, "Open government data catalogs: Current approaches and

- quality perspective," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 8061 LNCS, pp. 152–166, 2013.
- [166] M. Van der Waal, S. Węcel, K. Ermilov, I. Janev, V. Milošević, U., & Wainwright, *Linked Open Data - Creating Knowledge Out of Interlinked Data. Results of the LOD2 Project*. Cham: Springer International Publishing, 2014.
- [167] Y. K. Dwivedi *et al.*, "Driving innovation through big open linked data (BOLD): Exploring antecedents using interpretive structural modelling," *Inf. Syst. Front.*, vol. 19, no. 2, pp. 197–212, 2017.
- [168] F. Bauer and M. Kaltenböck, *Linked Open Data: A Quick Start Guide for Decision Makers*. 2012.
- [169] S. Foundation, "Ten Principles for Opening Up Government Information," *Sunlight Found.*, 2010.
- [170] K. Braunschweig, J. Eberius, M. Thiele, and W. Lehner, "The State of Open Data Limits of Current Open Data Platforms," *Proc. 21st World Wide Web Conf. 2012, Web Sci. Track WWW'12*, pp. 1–6, 2012.
- [171] M. Janssen, S. A. Chun, and J. R. Gil-Garcia, "Building the next generation of digital government infrastructures," *Gov. Inf. Q.*, vol. 26, no. 2, pp. 233–237, 2009.
- [172] T. Berners-lee, "Linked Data - Design Issues," *Design Issues*, 2006. .
- [173] T. Berners-lee, "Linked Data - Design Issues," *Des. Issues*, 2006.
- [174] N. Mihindikulasooriya, R. García-Castro, and M. Esteban-Gutiérrez, "Linked data platform as a novel approach for Enterprise Application Integration," *CEUR Workshop Proc.*, vol. 1034, pp. 1–12, 2013.
- [175] T. Heath and C. Bizer, *Linked Data Evelving the Web into a Global Data Space*, vol. 1(1). 2011.
- [176] M. Janssen and G. Kuk, "Big and Open Linked Data (BOLD) in research, policy, and practice," *J. Organ. Comput. Electron. Commer.*, vol. 26, no. 1–2, pp. 3–13, 2016.
- [177] B. Behkamal, M. Kahani, E. Bagheri, and Z. Jeremic, "A metrics-driven approach for quality assessment of linked open data," *J. Theor. Appl. Electron. Commer. Res.*, 2014.
- [178] Y. Demchenko, C. De Laat, and P. Membrey, "Defining architecture components of the Big Data Ecosystem," *2014 Int. Conf. Collab. Technol. Syst. CTS 2014*, pp. 104–112, 2014.
- [179] S. Trudgill, "Tansley, A.G. 1935: The use and abuse of vegetational concepts and terms. *Ecology* 16, 284–307," no. July, 2015.
- [180] A. Hein, J. Weking, M. Schrieck, M. Wiesche, M. Böhm, and H. Krcmar, "Value co-creation practices in business-to-business platform ecosystems," *Electron. Mark.*, vol. 29, no. 3, pp. 503–518, 2019.
- [181] M. I. S. Oliveira and B. F. Lóscio, "What is a data ecosystem?," *ACM Int. Conf. Proceeding Ser.*, 2018.
- [182] L. D. W. Thomas and E. Autio, "The processes of ecosystem emergence," *Acad. Manag. Proc.*, vol. 2015, no. 1, pp. 10453–10453, 2015.
- [183] S. Bresciani, A. Ferraris, M. Romano, and G. Santoro, "Digital Ecosystems," *Digit. Transform. Manag. Agil. Organ. A Compass to Sail Digit. World*, pp. 153–165, 2021.
- [184] M. Heimstädt, F. Saunderson, and T. Heath, "From Toddler to Teen: Growth of an Open Data Ecosystem," *JeDEM - eJournal eDemocracy Open Gov.*, vol. 6, no. 2, pp. 123–135, 2014.
- [185] F. De Prieelle, M. De Reuver, and J. Rezaei, "The Role of Ecosystem Data Governance in Adoption of Data Platforms by Internet-of-Things Data Providers: Case of Dutch Horticulture Industry," *IEEE Trans. Eng. Manag.*, pp. 1–11, 2020.
- [186] T. M. Harrison, T. A. Pardo, and M. Cook, "Creating Open Government Ecosystems: A Research and Development Agenda," pp. 900–928, 2012.
- [187] T. M. Harrison, "Open Data and Information Sharing in Developing Nations," pp. 311–313, 2014.
- [188] M. Heimstädt, F. Saunderson, and T. Heath, "Conceptualizing Open Data Ecosystems: A timeline analysis of Open Data development in the UK," vol. 1000, no. October 2013, 2014.
- [189] MIT, "MIT Total Data Quality Management Program and the International Conference on Information Quality." .
- [190] J. L. Soren Auer, Sebastian Tramp, Bert van Nu elen, Robert Isele, O. Lorenz B uhmman, Christian Dirschl, Pablo N. Mendes, Hugh Williams, and M. H. Erling, "Managing the Life-Cycle of Linked Data with the LOD2 Stack," vol. 7650, no. 00, pp. 362–374, 2012.
- [191] T. K. Das and P. M. Kumar, "BIG Data Analytics : A Framework for Unstructured Data Analysis,"

- vol. 5, no. 1, pp. 153–156, 2013.
- [192] B. M. Ferguson and I. B. Strategies, “Architecting A Platform For Big Data Analytics 2,” no. March, 2016.
- [193] P. Géczy, “BIG DATA MANAGEMENT: RELATIONAL FRAMEWORK,” vol. 6, no. 3, pp. 21–30, 2015.
- [194] world wide web Foundation, “OPEN DATA Barometer,” 2018.
- [195] D. Lee, “Building an Open Data Ecosystem – An Irish Experience,” in *ICEGOV '14: Proceedings of the 8th International Conference on Theory and Practice of Electronic Governance*, 2014, pp. 351–360.
- [196] G. Lakshen, V. Janev, and S. Vraneš, “Arabic linked drug dataset consolidating and publishing,” *Comput. Sci. Inf. Syst.*, vol. 18, no. 3, pp. 729–748, 2021.
- [197] P. Sawadogo and J. Darmont, “On data lake architectures and metadata management,” *J. Intell. Inf. Syst.*, 2020.
- [198] M. N. Mami, D. Graux, S. Scerri, H. Jabeen, S. Auer, and J. Lehmann, “Uniform access to multiform data lakes using semantic technologies,” in *ACM International Conference Proceeding Series*, 2019.
- [199] Z. M. Aljazzaf, “Modelling and measuring the quality of online services,” *Kuwait J. Sci.*, vol. 42, no. 3, pp. 134–157, 2015.
- [200] B. Talin, “What Is A Digital Ecosystem? – Understanding The Most Profitable Business Model,” *MoreThanDigital*. 2021.
- [201] C. Costa and M. Y. Santos, “Big Data: State-of-the-art concepts, techniques, technologies, modeling approaches and research challenges,” *IAENG Int. J. Comput. Sci.*, vol. 44, no. 3, pp. 285–301, 2017.
- [202] M. Acosta, A. Zaveri, E. Simperl, D. Kontokostas, S. Auer, and J. Lehmann, “Crowdsourcing linked data quality assessment,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 8219 LNCS, no. PART 2, pp. 260–276, 2013.
- [203] M. M. J. Ferney, L. Beltran Nicolas Estefan, and V. V. J. Alexander, “Assessing data quality in open data: A case study,” *2017 Congr. Int. Innov. y Tendencias en Ing. CONIITI 2017 - Conf. Proc.*, vol. 2018-Janua, pp. 1–5, 2018.
- [204] D. Kontokostas *et al.*, “Test-driven evaluation of Linked Data quality,” *WWW 2014 - Proc. 23rd Int. Conf. World Wide Web*, pp. 747–757, 2014.
- [205] A. Nikiforova and J. Bicevskis, “An extended data object-driven approach to data quality evaluation: Contextual data quality analysis,” *ICEIS 2019 - Proc. 21st Int. Conf. Enterp. Inf. Syst.*, vol. 1, no. Iceis, pp. 262–269, 2019.
- [206] A. Nikiforova, “Open Data Quality Evaluation: A Comparative Analysis of Open Data in Latvia,” *Balt. J. Mod. Comput.*, vol. 6, no. 4, pp. 363–386, 2018.
- [207] M. Yi, “Exploring the quality of government open data: Comparison study of the UK, the USA and Korea,” *Electronic Library*, vol. 37, no. 1. pp. 35–48, 2019.
- [208] M. Scannapieco, B. Pernici, and E. Pierce, “IP-UML: Towards a Methodology for Quality Improvement Based on the IP- Map Framework,” *Proc. Seventh Int. Conf. Inf. Qual. Spec.*, pp. 279–291, 2002.
- [209] S. Sarsfield, “How a Small Data Error Becomes a Big Data Quality Problem,” *DataInformed*. 2012.
- [210] E.-U. De Castilla, L. Mancha, M. Piattini, and E. D. C. La Mancha, “CALDEA : A Data Quality Model Based on Maturity Levels Ismael Caballero,” *Third Int. Conf. Qual. Software, 2003. Proceedings.*, 2003.
- [211] F. Sidi *et al.*, “Data Quality : A Survey of Data Quality Dimensions,” pp. 300–304, 2012.
- [212] G. K. Tayi and D. P. Ballou, “Examining Data Quality,” *Commun. ACM*, vol. 41, no. 2, pp. 54–57, 1998.
- [213] Y. W. Lee and L. L. Pipino, *Journey to Data Quality*. The MIT Press, 2002.
- [214] M. Amadeo and C. Campolo, “Multi-Source Data Retrieval in IoT via Named Data Networking,” *Zeitschrift f??r Phys. D Atoms, Mol. Clust.*, vol. 26, no. 1, pp. 246–248, 1993.
- [215] J. Debattista, C. Lange, and S. Auer, “Representing dataset quality metadata using multi-dimensional views,” in *ACM International Conference Proceeding Series*, 2014.
- [216] C. Cappiello, C. Francalanci, and B. Pernici, “Data quality assessment from the user’s perspective,” *Proc. ACM SIGMOD Int. Conf. Manag. Data*, pp. 68–73, 2004.
- [217] A. Sinaeepourfard, X. Masip-Bruin, J. Garcia, and E. Marín-Tordera, “A Survey on Data Lifecycle

Models: Discussions toward the 6Vs Challenges,” *Tech. Rep.*, 2015.

- [218] J. Debattista, S. Auer, and C. Lange, “Luzzu - A methodology and framework for linked data quality assessment,” *J. Data Inf. Qual.*, 2016.
- [219] N. Mihindikulasooriya, G. Rizzo, R. Troncy, O. Corcho, and R. García-Castro, “A two-fold quality assurance approach for dynamic knowledge bases: The 3cixty use case,” in *CEUR Workshop Proceedings*, 2016.
- [220] S. Neumaier, “Open data quality Assessment and Evolution of (Meta-)Data Quality in the Open Data Landscape,” 2015.
- [221] R. Máchová and M. Lněnička, “Evaluating the quality of open data portals on the national level,” *J. Theor. Appl. Electron. Commer. Res.*, vol. 12, no. 1, 2017.
- [222] I. Taleb, H. T. E. Kassabi, M. A. Serhani, R. Dssouli, and C. Bouhaddioui, “Big Data Quality: A Quality Dimensions Evaluation,” in *Proceedings - 13th IEEE International Conference on Ubiquitous Intelligence and Computing, 13th IEEE International Conference on Advanced and Trusted Computing, 16th IEEE International Conference on Scalable Computing and Communications, IEEE Internationala*, 2017.
- [223] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, “Methodologies for data quality assessment and improvement,” *ACM Comput. Surv.*, vol. 41, no. 3, 2009.
- [224] T. C. Redman and B. A. Godfrey, *Data Quality for the Information Age*. 1996.
- [225] D. M. Strong, Y. W. Lee, and R. Y. Wang, “Data quality in context,” *Commun. ACM*, 1997.
- [226] L. P. English, “The TIQM @ Quality System for Total Information Quality Management,” *M.I.T. Ind. Symp.*, pp. 67–86, 2009.
- [227] K.-T. Huang, R. Y. Wang, and Y. W. Lee, *Quality Information and Knowledge*. 1998.
- [228] C. Bizer, “Quality-Driven Information Filtering in the Context of Web-Based Information Systems,” 2007.
- [229] A. Zaveri, “Linked Data Quality Assessment and its Application to Societal Progress Measurement,” no. November 1984, pp. 1–158, 2015.
- [230] J. Umbrich, N. Sebastian, and A. Polleres, “Towards assessing the quality evolution of Open Data portals,” *Futur. Internet Things Cloud (FiCloud), 2015 3rd Int. Conf.*, pp. 404–411, 2015.
- [231] M. Song, K. Liu, R. Abromitis, and T. L. Schleyer, “Reusing electronic patient data for dental clinical research: A review of current status,” *Journal of Dentistry*. 2013.
- [232] I. Danciu *et al.*, “Secondary use of clinical data: The Vanderbilt approach,” *J. Biomed. Inform.*, 2014.
- [233] D. I. *et al.*, “Secondary use of clinical data: The Vanderbilt approach,” *J. Biomed. Inform.*, 2014.
- [234] T. Botsis, G. Hartvigsen, F. Chen, and C. Weng, “Secondary Use of EHR: Data Quality Issues and Informatics Opportunities,” *AMIA Jt. Summits Transl. Sci. proceedings. AMIA Jt. Summits Transl. Sci.*, 2010.
- [235] R. S. Mans, W. M. P. Van Der Aalst, and R. J. B. Vanwersch, *Process Mining in Healthcare Evaluating and Exploiting Operational Healthcare Processes*. 2015.
- [236] R. P. J. C. Bose, R. S. Mans, and W. M. P. Van Der Aalst, “Wanna improve process mining results?,” in *Proceedings of the 2013 IEEE Symposium on Computational Intelligence and Data Mining, CIDM 2013 - 2013 IEEE Symposium Series on Computational Intelligence, SSCI 2013*, 2013.
- [237] I. Abaker *et al.*, “The rise of ‘ big data ’ on cloud computing : Review and open research issues,” *Inf. Syst.*, vol. 47, pp. 98–115, 2015.
- [238] Kitchin Rob, “The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences (2014),” *J. Reg. Sci.*, vol. 56, no. 4, pp. 722–723, 2014.
- [239] S. LaValle, E. Lesser, R. Shockley, M. S. Hopkins, and N. Kruschwitz, “LaValle, Steve, et al. 2011 Big Data, analytics and the path from insights to value,” *MIT Sloan Manag. Rev.*, vol. 52, no. 2, pp. 21–31, 2011.
- [240] T. Jetzek, “Innovation in the Open Data Ecosystem : Exploring the role of real options thinking and multi-sided platforms for sustainable value generation through open data Innovation in the Open Data Ecosystem Exploring the role of real options thinking and multi - ,” no. February 2017, 2018.
- [241] B. and Q. Chen, D. Asaolu, “Big Data Analytics In The Public Sector_ A Case Study Of NEET Analysis

For The London Boroughs,” in *IADIS International Journal on Computer Science & Information Systems*, 2016, p. 11(2).

- [242] P. Colpaert, P. Mechant, E. Mannens, and R. Van De Walle, “The 5 stars of open data portals,” no. January, 2013.
- [243] J. E. Ross, *Total quality management: Text, cases, and readings: Third edition*. 2017.
- [244] R. Singh and K. Singh, “A Descriptive Classification of Causes of Data Quality Problems in Data Warehousing,” *Int. J. Comput. Sci. Issues*, vol. 7, no. 3, pp. 41–50, 2010.
- [245] E. Rahm and H. Do, “Data cleaning: Problems and current approaches,” *IEEE Data Eng. Bull.*, 2000.
- [246] S. E. Madnick, R. Y. Wang, and Y. W. Lee, “Overview and Framework for Data and Information Quality Research,” vol. 1, no. 1, pp. 1–22, 2009.
- [247] M. Ge and M. Helfert, “A REVIEW OF INFORMATION QUALITY RESEARCH,” in *International Conference on Information Quality*, 2007.
- [248] Y. Wand and R. Y. Wang, “Anchoring data quality dimensions in ontological foundations,” *Commun. ACM*, vol. 39, pp. 86–95, 1996.
- [249] M. Bovee, R. P. Srivastava, and B. Mak, “A conceptual framework and belief-function approach to assessing overall information quality,” in *International Journal of Intelligent Systems*, 2003.
- [250] Naumann Felix, *Quality-Driven Query Answering for Integrated Information Systems*, vol. 2261. 2002.
- [251] scannapieca monica batini Carlo, *Data quality concepts, methodologies, and techniques*. Springer Berlin Heidelberg, 2016.
- [252] P. Woodall, A. Borek, and A. K. Parlikad, “Data quality assessment: The Hybrid Approach,” *Inf. Manag.*, 2013.
- [253] P. Woodall, A. Borek, J. Gao, M. Oberhofer, and A. Koronios, “An investigation of how data quality is affected by dataset size in the context of big data analytics,” in *Proceedings of the 19th International Conference on Information Quality, ICIQ 2014*, 2014.
- [254] B. T. Hazen, C. A. Boone, J. D. Ezell, and L. A. Jones-Farmer, “Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications,” *Int. J. Prod. Econ.*, 2014.
- [255] O. Kwon, N. Lee, and B. Shin, “Data quality management, data usage experience and acquisition intention of big data analytics,” *Int. J. Inf. Manage.*, 2014.
- [256] D. Rao, V. N. Gudivada, and V. V. Raghavan, “Data quality issues in big data,” in *Proceedings - 2015 IEEE International Conference on Big Data, IEEE Big Data 2015*, 2015.
- [257] M. A. Serhani, H. T. El Kassabi, I. Taleb, and A. Nujum, “An hybrid approach to quality evaluation across big data value chain,” in *Proceedings - 2016 IEEE International Congress on Big Data, BigData Congress 2016*, 2016.
- [258] I. Taleb and M. A. Serhani, “Big Data Pre-Processing: Closing the Data Quality Enforcement Loop,” in *Proceedings - 2017 IEEE 6th International Congress on Big Data, BigData Congress 2017*, 2017.
- [259] C. Xie, J. Gao, and C. Tao, “Big data validation case study,” in *Proceedings - 3rd IEEE International Conference on Big Data Computing Service and Applications, BigDataService 2017*, 2017.
- [260] T. Catarci, M. Scannapieco, M. Console, and C. Demetrescu, “My (fair) big data,” in *Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017*, 2017.
- [261] I. Taleb, M. A. Serhani, and R. Dssouli, “Big Data Quality: A Survey,” in *Proceedings - 2018 IEEE International Congress on Big Data, BigData Congress 2018 - Part of the 2018 IEEE World Congress on Services*, 2018.
- [262] D. Ardagna, C. Cappiello, W. Samá, and M. Vitali, “Context-aware data quality assessment for big data,” *Futur. Gener. Comput. Syst.*, 2018.
- [263] I. El Alaoui, Y. Gahi, and R. Messoussi, “Big data quality metrics for sentiment analysis approaches,” in *ACM International Conference Proceeding Series*, 2019.
- [264] J. L. Kulikowski, “Data quality assessment,” *Encycl. Database Technol. Appl.*, vol. 45, no. 4, pp. 116–120, 2005.
- [265] L. L. Pipino, Y. W. Lee, and R. Y. Wang, “Data Quality Assessment,” vol. 45, no. 4, pp. 211–218, 2002.

- [266] P. Woodall, M. Oberhofer, and A. Borek, "A classification of data quality assessment and improvement methods," *Int. J. Inf. Qual.*, vol. 3, no. 4, pp. 298–321, 2014.
- [267] P. Hitzler *et al.*, "Quality Assessment for Linked Data: A Survey A Systematic Literature Review and Conceptual Framework," *Semant. Web*, vol. 1, pp. 1–5, 2012.
- [268] C. Batini and P. Milano, "Methodologies for data quality assessment and improvement," vol. 41, no. 3, 2009.
- [269] R. Price and G. Shanks, "A Semiotic Information Quality Framework: Development and Comparative Analysis," *Enacting Res. Methods Inf. Syst.*, pp. 219–250, 2016.
- [270] C. Fürber and M. Hepp, "Towards a vocabulary for data quality management in semantic web architectures," *ACM Int. Conf. Proceeding Ser.*, pp. 1–8, 2011.
- [271] Data Management Association (DAMA)/ UK Working and Group, "THE SIX PRIMARY DIMENSIONS FOR DATA." p. 17, 2013.
- [272] D. Boyd and K. Crawford, "Critical questions for big data - Provocations for a cultural, technological, and scholarly phenomenon," *Inf. Tarsad.*, no. 2, pp. 7–23, 2012.
- [273] D. McGilvray, "Ten Steps to Quality Data and Trusted Information™," *MIT Inf. Qual. Ind. Symp.*, no. Enterprise 2008, pp. 2–19, 2009.
- [274] Y. W. Lee, "Crafting Rules: Context-Reflective Data Quality Problem Solving," *J. Manag. Inf. Syst.*, vol. 20, no. 3, pp. 93–119, 2003.
- [275] M. Knuth, D. Kontokostas, and H. Sack, "Linked data quality: Identifying and tackling the key challenges," in *CEUR Workshop Proceedings*, 2014.
- [276] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, and S. Auer, "Quality Assessment Methodologies for Linked Open Data: A Systematic Literature Review and Conceptual Framework," *Semant. Web – Interoperability, Usability, Appl.*, vol. 1, p. 33, 2012.
- [277] A. Zaveri *et al.*, "User-driven quality evaluation of DBpedia," *ACM Int. Conf. Proceeding Ser.*, pp. 97–104, 2013.
- [278] F. Radulovic, N. Mihindukulasooriya, R. García-Castro, and A. Gómez-Pérez, "A comprehensive quality model for Linked Data," *Semant. Web*, vol. 9, no. 1, pp. 3–24, 2018.
- [279] D. Loshin, *Enterprise Knowledge Management: The Data Quality Approach*. 2001.
- [280] V. Peralta, "Data Freshness and Data Accuracy: A State of the Art," *Inst. Comput. Fac. Ing. Univ. la Repub.*, 2006.
- [281] C. Fürber and M. Hepp, "SWIQA - A Semantic Web information quality assessment framework," in *19th European Conference on Information Systems, ECIS 2011*, 2011.
- [282] F. Salgé, "Semantic accuracy," in *Elements of Spatial Data Quality*, 1995.
- [283] J. Lehmann, D. Gerber, M. Morsey, and A. Ngonga, "DeFacto - Deep Fact Validation," 2012.
- [284] Y. Lei *et al.*, "A Framework for Evaluating Semantic Metadata," pp. 135–142, 2007.
- [285] D. Loshin, "Monitoring Data Quality Performance Using Data Quality Metrics," *Informatica*, 2006.
- [286] B. Gatling, S. Champlin, B. Weigel, and E. Dharwad, *Enterprise Information Management with SAP*. Galileo Press, 2012.
- [287] Health Information and Quality Authority, "International Review of Data Quality," no. April, pp. 1–59, 2011.
- [288] F. Maali, R. Cyganiak, and V. Peristeras, "Enabling interoperability of government data catalogues," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2010.
- [289] J. Kučera, D. Chlapek, and M. Nečaský, "Open Government Data Catalogs: Current Approaches and Quality Perspective," 2013.
- [290] E. Sovey and P. Petocz, "Quality in statistics," *A Panor. Stat.*, no. May, pp. 157–164, 2016.
- [291] S. L. C. Stvilia, B. Gasser, L. Twidale M., B., "A framework for information quality assessment," *JASIST*, vol. 58, no. 12, pp. 1720–1733, 2007.
- [292] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web," *World Wide Web Internet Web Inf. Syst.*, 1998.
- [293] C. S. Bond, "Web users' information retrieval methods and skills," *Online Inf. Rev.*, 2004.
- [294] D. Ballou, R. Wang, H. Pazer, and G. K. Tayi, "Modeling information manufacturing systems to

- determine information product quality,” *Manage. Sci.*, vol. 44, no. 4, pp. 462–484, 1998.
- [295] R. Y. Wang, “A product perspective on total data quality management,” *Commun. ACM*, vol. 41, no. 2, pp. 58–65, 1998.
- [296] R. Price, D. Neiger, and G. Shanks, “Developing a Measurement Instrument for Subjective Aspects of Information Quality,” *Commun. Assoc. Inf. Syst.*, vol. 22, 2008.
- [297] B. Heinrich, M. Kaiser, and M. Klier, “How To Measure Data Quality?,” in *Proceedings of the 16th International Conference on Information Quality (ICIQ-2011)*, 2011.
- [298] M. Kaiser, B. Heinrich, and M. Kaiser, “Metrics for measuring data quality – Foundations for an economic data quality management Diskussionspapier WI-199 Metrics for measuring data quality – Foundations for an economic data quality management von Beitrag für: 2nd International Conference on So,” no. May, 2014.
- [299] J. P. Hirdes *et al.*, “An evaluation of data quality in Canada’s continuing care reporting system (CCRS): Secondary analyses of Ontario data submitted between 1996 and 2011,” *BMC Med. Inform. Decis. Mak.*, 2013.
- [300] D. P. Ballou and H. L. Pazer, “Modeling Data and Process Quality in Multi-Input, Multi-Output Information Systems,” *Manage. Sci.*, vol. 31, no. 2, pp. 150–162, 1985.
- [301] M. Gamble and C. Goble, “Quality, trust, and utility of scientific data on the web: Towards a joint model,” in *Proceedings of the 3rd International Web Science Conference, WebSci 2011*, 2011.
- [302] J. Golbeck and A. Mannes, “Using trust and provenance for content filtering on the semantic web,” in *CEUR Workshop Proceedings*, 2006.
- [303] Y. Gil and D. Artz, “Towards content trust of web resources,” *Web Semant.*, 2007.
- [304] S. Shekarpour and S. D. Katebi, “Modeling and evaluation of trust with an extension in semantic web,” *J. Web Semant.*, 2010.
- [305] A. Hogan, J. Umbrich, A. Harth, R. Cyganiak, A. Polleres, and S. Decker, “An empirical survey of Linked Data conformance,” *J. Web Semant.*, 2012.
- [306] A. Rula, M. Palmonari, and A. Maurino, “Capturing the age of linked open data: Towards a dataset-independent framework,” in *Proceedings - IEEE 6th International Conference on Semantic Computing, ICSC 2012*, 2012.
- [307] M. Ben Ellefi *et al.*, “RDF dataset profiling – a survey of features, methods, vocabularies and applications,” *Semant. Web*, 2018.
- [308] A. Rula and A. Zaveri, “Methodology for assessment of linked data quality,” in *CEUR Workshop Proceedings*, 2014.
- [309] V. R. Basili, G. Caldiera, and H. D. Rombach, “The goal question metric approach,” *Encycl. Softw. Eng.*, 1994.
- [310] E. Ruckhaus, M. E. Vidal, S. Castillo, O. Burguillos, and O. Baldizan, “Analyzing linked data quality with LiQuate,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, 2014.
- [311] D. Kontokostas, A. Zaveri, S. Auer, and J. Lehmann, “Triplecheckmate: A tool for crowdsourcing the quality assessment of linked data,” in *Communications in Computer and Information Science*, 2013.
- [312] C. Guéret, P. Groth, C. Stadler, and J. Lehmann, “Assessing linked data mappings using network measures,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2012.
- [313] P. N. Mendes, H. Mühleisen, and C. Bizer, “Sieve: Linked Data quality assessment and fusion,” *ACM Int. Conf. Proceeding Ser.*, no. March, pp. 116–123, 2012.
- [314] T. Knap *et al.*, “ODCleanStore: A framework for managing and providing integrated linked data on the web,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2012.
- [315] B. Spahiu, “Profiling the Linked (Open) Data,” in *CEUR Workshop Proceedings*, 2015.
- [316] J. Golbeck, B. Parsia, and J. Hendler, “Trust networks on the Semantic Web,” in *Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science)*, 2003.
- [317] K. C. Feeney, D. O’Sullivan, W. Tai, and R. Brennan, “Improving curated web- data quality with

- structured harvesting and assessment," *Int. J. Semant. Web Inf. Syst.*, 2014.
- [318] C. Böhm *et al.*, "Profiling linked open data with ProLOD," in *Proceedings - International Conference on Data Engineering*, 2010.
- [319] O. Hartig, "Trustworthiness of Data on the Web," *Science (80-.)*, 2008.
- [320] Y. Gil and V. Ratnakar, "Trusting information sources one citizen at a time," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2002.
- [321] B. Hyland, G. Ateazing, and B. Villazón-Terrazas, "Best Practices for Publishing Linked Data (www.w3.org/TR/ld-bp/)." .
- [322] R. Verborgh and J. De Roo, "Drawing conclusions from linked data on the web: The EYE reasoner," *IEEE Softw.*, 2015.
- [323] G. Ateazing *et al.*, "Publishing Linked Data Requires More than Just Using a Tool," in *W3C 2013, Workshop on Open Data on the Web*, 2013, no. April, pp. 1–5.
- [324] V. Janev, V. Mijovic, and S. Vraneš, "Using the linked data approach in European e-government systems: Example from Serbia," *Int. J. Semant. Web Inf. Syst.*, vol. 14, no. 2, pp. 27–46, 2018.
- [325] B. Hyland and D. Wood, "Linking Government Data," *Link. Gov. Data*, pp. 3–26, 2011.
- [326] M. Hausenblas, "Linked data life cycles," *Linked Data Lifecycles*, 2011. .
- [327] B. Villazón-Terrazas, L. M. Vilches-Blázquez, O. Corcho, and A. Gómez-Pérez, "Methodological Guidelines for Publishing Government Linked Data," *Link. Gov. Data*, pp. 27–49, 2011.
- [328] S. Auer *et al.*, "Managing the life-cycle of linked data with the LOD2 stack," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2012.
- [329] G. Lakshen, V. Janev, and S. Vraneš, "Linking Open Drug Data : The Arabic dataset," pp. 3–7, 2019.
- [330] C. Buil-Aranda, A. Hogan, J. Umbrich, and P. Y. Vandenbussche, "SPARQL web-querying infrastructure: Ready for action?," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 8219 LNCS, no. PART 2, pp. 277–293, 2013.
- [331] A. Hadhiatma, "Improving data quality in the linked open data: A survey," *J. Phys. Conf. Ser.*, vol. 978, no. 1, 2018.
- [332] G. Weber, "The World's 10 Most Influential Languages," *Lang. Today*, 3, 1997.
- [333] A. M. Al-Zoghby, A. S. E. Ahmed, and T. T. Hamza, "Arabic semantic web applications - A survey," *J. Emerg. Technol. Web Intell.*, vol. 5, no. 1, pp. 52–69, 2013.
- [334] S. R. El-Beltagy, M. Hazman, and A. Rafea, "Ontology based annotation of text segments," *Proc. ACM Symp. Appl. Comput.*, pp. 1362–1367, 2007.
- [335] الدكتور كادان الجمعة و ثائر محمد, "تجزئة الاهداف باستخدام أنطولوجيا الافعال العربية في البحث عن خدمات الوب الموجهة بالاهداف," *Tishreen Univ. J. Res. Sci. Stud. - Eng. Sci. Ser.*, vol. 4, no. 37, pp. 721–737, 2015.
- [336] K. Rodgers and N. Massac, "Misinformation: A Threat to the Public's Health and the Public Health System," *Journal of Public Health Management and Practice*. 2020.
- [337] A. Jentzsch, M. Samwald, and B. Andersson, "Linking Open Drug Data," *I-Semantics '09 Proc. Int. Conf. Semant. Syst.*, 2009.
- [338] O. Bodenreider, "The Unified Medical Language System (UMLS): Integrating biomedical terminology," *Nucleic Acids Res.*, vol. 32, no. DATABASE ISS., 2004.
- [339] M. Kuhn, C. von Mering, M. Campillos, L. J. Jensen, and P. Bork, "STITCH: Interaction networks of chemicals and proteins," *Nucleic Acids Res.*, vol. 36, no. SUPPL. 1, pp. 684–688, 2008.
- [340] J. Lathem and J. Lathem, "SA-REST: Semantically Interoperable and Easier-," vol. 11, no. December, pp. 91–94, 2007.
- [341] J. Volz, C. Bizer, M. Gaedke, and G. Kobilarov, "Silk - A Link Discovery Framework for the Web of Data," *CEUR Workshop Proc.*, vol. 538, 2009.
- [342] O. Hassanzadeh, A. Kementsietsidis, L. Lim, R. J. Miller, and M. Wang, "A framework for semantic link discovery over relational data," *Int. Conf. Inf. Knowl. Manag. Proc.*, no. January, pp. 1027–1036, 2009.
- [343] M. Samwald *et al.*, "Linked Open drug data for pharmaceutical research and development," *J. Cheminform.*, vol. 3, no. 5, pp. 1–6, 2011.

- [344] M. Jovanovik, "Linked Data Application Development Methodology Linked Data Application Development Methodology Executive Summary of the PhD Thesis," no. November 2016, 2017.
- [345] F. Belleau, M. A. Nolin, N. Tourigny, P. Rigault, and J. Morissette, "Bio2RDF: Towards a mashup to build bioinformatics knowledge systems," *J. Biomed. Inform.*, vol. 41, no. 5, pp. 706–716, 2008.
- [346] B. Chen *et al.*, "Chem2Bio2RDF: A semantic framework for linking and data mining chemogenomic and systems chemical biology data," *BMC Bioinformatics*, vol. 11, 2010.
- [347] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "DBpedia: A nucleus for a Web of open data," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2007.
- [348] G. A. Lakshen, V. Janev, and S. Vraneš, "Challenges in quality assessment of Arabic dbpedia," *ACM Int. Conf. Proceeding Ser.*, 2018.
- [349] G. Lakshen, V. Janev, and S. Vraneš, *Linking open drug data: Lessons learned*, vol. 11703 LNCS. 2019.
- [350] V. Mijović, V. Janev, D. Paunović, and S. Vraneš, "Exploratory spatio-temporal analysis of linked statistical data," *J. Web Semant.*, vol. 41, pp. 1–8, 2016.
- [351] V. Kundra, "Digital Fuel of the 21st Century: Innovation through Open Data and the Network Effect," *Public Policy*, no. January, p. 21, 2012.

Biography

Guma Abdulkhader Lakshen was born on 1st January 1963 in a small town called Al-Asabaa- Tripoli, Libya. Finished his elementary, preparatory, secondary schooling in his town (1969-1982). He was known for his determination and hardworking at school and for learning in general. Guma was chosen among other students from all over Libya for a scholarship in 1983 to study in England by the newly found Libyan Iron and steel company LISCO. Guma studied for his Bachelor's Degree, at the Computer department, college of Cardiff, University of Wales, 1985-1989, Guma gained a 2nd class honors degree.

Guma started his career by working at LISCO from 1989-1997, he mainly worked as a programmer, head of the computer section, and computer department manager, Guma was the Libyan team leader working with Dastur Engineering Company (Indian) a contractor with LISCO for establishing a production, maintenance management, and administrative systems.

In July 1997, Guma started at the Libyan Pharmaceutical/medical Company GMPMSCO as computer department manager from 1997- 2000. Guma was a manager of the computer and information manager, Al-Maya medical factory, GMPMSCO, Al-Maya, Libya (2000-2002).

In September 2002, Guma was nominated to do a Master's Study in computer systems at the Libyan Higher academy, Libya; graduated in 2006 with (3.79/4.00).

Guma was appointed as a manager of Al-rabta Pharmaceutical factory, GMPMSCO Libya, Al-rabta, Libya (2007-2009). Guma was Logistics manager (2009-2011), Libyan Pharmaceutical/medical Company, GMPMSCO, Tripoli, Libya.

Guma worked as a Data Analyst (part-time) at Almadena media and information center, Ministry of information and telecommunications (2006-2011), Tripoli, Libya.

Guma moved to work as a lecturer and as head of the computer section at Head of Computer section, Zintan teaching college, Al-Jabal Al-Gharbe University, Zentan, Libya (2011-2013).

Guma was granted a scholarship in 2013 to do doctoral study in the Republic of Serbia.

During the years from 1989 till 2013 Guma was involved in these notable projects among others:

1. Maintenance Management System MMS, a computerized maintenance management system, LISCO. 1989-1993.
2. Marketing Management Application. MMA, sales distribution system based on geographical and inhabitation of Libya, GMPMSCO, 2001-2003.
3. Designing HR management system, GMPMSCO, 2002-2005.
4. Classifying, monitoring, surveillance, of production equipment and activities at Al-rabta Pharmaceutical factory, as required by OPCW. 2002-2011.
5. Study the requirements of unifying a computerized system at Al-Jabal Al-Gharbe University, 2011-2012.

образац изјаве о ауторству

Изјава о ауторству

Име и презиме аутора _____ Гума ЛАКШЕН (Guma Abdulkhader LAKSHEN) _____

Број индекса _____ 2013/5054 _____

Изјављујем

да је докторска дисертација под насловом

Окружење за анализу и оцену квалитета великих и повезаних података

- резултат сопственог истраживачког рада;
- да дисертација у целини ни у деловима није била предложена за стицање друге дипломе према студијским програмима других високошколских установа;
- да су резултати коректно наведени и
- да нисам кршио/ла ауторска права и користио/ла интелектуалну својину других лица.

Потпис аутора



У Београду, _____ 22/11/2021 _____

Изјава о истоветности штампане и електронске верзије докторског Рада

Име и презиме аутора _____ Гума ЛАКШЕН (Guma Abdulkhader LAKSHEN) _____

Број индекса _____ 2013/5054 _____

Студијски програм _____ Докторске академске студије _____

Наслов рада _____ Окружење за анализу и оцену квалитета великих и повезаних података

Ментор _____ Професор Др Сања Вранеш _____

Изјављујем да је штампана верзија мог докторског рада истоветна електронској верзији коју сам предао/ла ради похрањивања у **Дигиталном репозиторијуму Универзитета у Београду**.

Дозвољавам да се објаве моји лични подаци везани за добијање академског назива доктора наука, као што су име и презиме, година и место рођења и датум одбране рада. Ови лични подаци могу се објавити на мрежним страницама дигиталне библиотеке, у електронском каталогу и у публикацијама Универзитета у Београду.

Потпис аутора



У Београду, _____ 22/11/2021 _____

Изјава о коришћењу

Овлашћујем Универзитетску библиотеку „Светозар Марковић“ да у Дигитални репозиторијум Универзитета у Београду унесе моју докторску дисертацију под насловом:

Окружење за анализу и оцену квалитета великих и повезаних података

која је моје ауторско дело.

Дисертацију са свим прилозима предао/ла сам у електронском формату погодном за трајно архивирање.

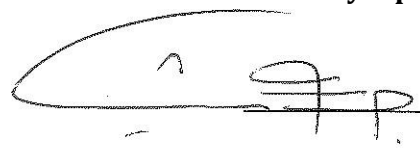
Моју докторску дисертацију похрањену у Дигиталном репозиторијуму Универзитета у Београду и доступну у отвореном приступу могу да користе сви који поштују одредбе садржане у одабраном типу лиценце Креативне заједнице (Creative Commons) за коју

сам се одлучио/ла.

1. Ауторство (CC BY)
2. Ауторство – некомерцијално (CC BY-NC)
3. Ауторство – некомерцијално – без прерада (CC BY-NC-ND)
4. Ауторство – некомерцијално – делити под истим условима (CC BY-NC-SA)
5. Ауторство – без прерада (CC BY-ND)
6. Ауторство – делити под истим условима (CC BY-SA)

(Молимо да заокружите само једну од шест понуђених лиценци.
Кратак опис лиценци је саставни део ове изјаве).

Потпис аутора



У Београду, _____22/11/2021_____