

AMERICAN UNIVERSITY OF BEIRUT

KGFUSIONX: LINKING, COMBINING, AND EXPLORING  
DATA THROUGH KNOWLEDGE GRAPHS

by  
SHADI ABDALLAH YOUSSEF

A thesis  
submitted in partial fulfillment of the requirements.  
for the degree of Master of Science in Business Analytics  
to the Suliman S. Olayan School of Business  
at the American University of Beirut

Beirut, Lebanon  
January 2024

AMERICAN UNIVERSITY OF BEIRUT

KGFUSIONX: LINKING, COMBINING, AND EXPLORING  
DATA THROUGH KNOWLEDGE GRAPHS

by  
SHADI ABDALLAH YOUSSEF

Approved by:

Signature



---

Dr. Fouad Zablith, Associate Professor -Director  
Suliman S. Olayan School of Business

Advisor

Signature

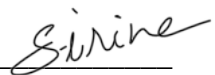


---

Dr. Bijan Azad, Professor  
Suliman S. Olayan School of Business

Co-Advisor

Signature



---

Dr. Sirine Taleb, Visiting Assistant Professor  
Suliman S. Olayan School of Business

Member of Committee

Signature



---

Dr. Walid Nasr, Associate Professor  
Suliman S. Olayan School of Business

Member of Committee

Date of thesis defense: January 26, 2024


# AMERICAN UNIVERSITY OF BEIRUT

## THESIS RELEASE FORM

Student Name: Youssef Shadi Abdallah  
Last First Middle

I authorize the American University of Beirut, to: (a) reproduce hard or electronic copies of my thesis; (b) include such copies in the archives and digital repositories of the University; and (c) make freely available such copies to third parties for research or educational purposes:

- As of the date of submission
- One year from the date of submission of my thesis.
- Two years from the date of submission of my thesis.
- Three years from the date of submission of my thesis.

 February 6, 2024  
Signature Date

## ACKNOWLEDGEMENTS

I would like to start with my deeply grateful to my thesis advisor, Dr Fouad Zablith, for his invaluable guidance, unlimited support, and expertise throughout this research. His comments and guidance are a vital part of the success of this thesis from research to writing. I cannot expect any advisor better than him for my thesis. I also extend my thanks to the second reader, Dr Bijan Azad, for his constructive comments and feedback throughout the thesis journey.

To my family, friends, and fellow students, your encouragement and camaraderie have been a constant source of strength and inspiration. I appreciate the participants who generously contributed their time and insights, without whom this research would not have been possible. The completion of this master's thesis marks a significant milestone in my academic journey, and I am thankful for the contributions of all those who have been part of this achievement.

# ABSTRACT

## OF THE THESIS OF

Shadi Abdallah Youssef for Master of Science in Business Analytics  
Major: Business Analytics

Title: KGFusionX: Linking, Combining, and Exploring Data Through Knowledge Graphs

In the realm of data exploration, the persistent challenges of data disconnection and inconsistency often hinder the efficiency of data analysts, especially in terms of data enrichment and aggregation. This thesis focuses on addressing the following research questions: How can we improve data integration and reuse of data in a clean and downloadable format to facilitate data analysis? Moreover, how can we contextually expand data on the fly to leverage its value and enhance data exploration? This work proposes KGFusionX, a knowledge graph centered framework that recognizes the time-intensive nature of data enrichment and integration. The study employs a backend implementation utilizing knowledge graphs to seamlessly connect disparate datasets. Several datasets from Lebanon covering different domains (e.g. health care, economy, education, and others) were converted and published as openly accessible knowledge graphs in a triple store repository (749,500 triples). This conversion allows efficient and fast aggregation of data because of the connections generated by knowledge graphs. Also, it is integrated with open linked data sources that serves as a resource to expand the data. The framework is showcased through an online platform built with Streamlit that allows users to select, combine, and download tabular data that can be used in other visualization exploration tools (e.g. PowerBI and Tableau). The approach was evaluated by data analysts and two use cases. Potential pickup of our platform was expressed by users who relied on the tool to analyze school and university challenges in rural areas, in addition to boosting tourism in Lebanon. The results demonstrated a significant improvement in data exploration efficiency, and better visuals with the knowledge graph-driven approach proving successful in overcoming the challenges posed by disconnection, inconsistency, and enrichment. This research primarily contributes to streamlining data exploration using the high potential of knowledge graphs to support data aggregation, data enrichment and visual data analysis.

## TABLE OF CONTENTS

ACKNOWLEDGEMENTS .....	1
ABSTRACT.....	2
ILLUSTRATIONS .....	5
TABLES .....	6
ABBREVIATIONS.....	7
INTRODUCTION.....	8
1.1 Aims and Objectives.....	11
1.2 Thesis Outline.....	11
LITERATURE REVIEW .....	13
2.1 Introduction .....	13
2.2 Frameworks for Knowledge graphs generation .....	13
2.3 Frameworks that use Semantic web in data exploration.....	15
2.4 Data Enrichment using Knowledge graphs.....	25
2.5 Gap and proposed Solution .....	26
RESEARCH METHODOLOGY .....	27
3.1 Data Sources:.....	28
3.2 Data Preparation .....	30
3.2.1 HTML TO CSV .....	31

3.2.2 PDF TO CSV .....	32
3.3 RDF Generator: CSV to RDF .....	33
3.4 Storing RDFs in a Triple Store:.....	39
3.5 Presenting the data to the user: “KGFusionX.”.....	41
3.6 Data Integration .....	44
3.7 Contextual Data Enrichment .....	45
3.8 Data Download.....	46
<b>VALIDATION AND RESULTS .....</b>	<b>47</b>
4.1. Boosting Tourism in Lebanon.....	47
4.2. Education in Rural Areas .....	50
<b>CONCLUSION .....</b>	<b>53</b>
5.1 Meeting the aims and objectives .....	54
5.2 Research Limitations .....	55
5.3 Future Work .....	55
<b>APPENDIX.....</b>	<b>57</b>
<b>REFERENCES.....</b>	<b>58</b>

# ILLUSTRATIONS

## Figure

1. Research Methodology Framework (high level) .....	27
2. An HTML page .....	29
3. Ontology Schema .....	36
4. Sample RDF Observation .....	37
5. “KGFusionX Interface” .....	41
6. Choosing your dataset and filtering functionalities .....	43
7. Merging of Typhus and Viral Hepatitis datasets .....	44
8. Illustration on Contextual Data Enrichment.....	45
9. Data Enrichment Functionality.....	46
10. Status of internet availability and phone network in the Lebanese Districts and Towns.....	49
11. Dashboard of project representing Public Transportation and Hotels in Lebanon by district.....	50
12. Number of Public schools per citizen .....	51
13. Percentage of Elderly in Zahle villages .....	52



# TABLES

## Table

1. Representing the datasets used, their sources, and number of triples generated.....38

## ABBREVIATIONS

RDF: Resource Description Framework

RML: RDF Mapping Language

EDL: Electricity of Lebanon

MOPH: Ministry of Public Health

SQL: Structured Query Language

KG: Knowledge Graph

# CHAPTER 1

## INTRODUCTION

In recent years, the generation of massive amounts of data has become increasingly prevalent across various industries. However, a significant challenge persists as much of this data remains unorganized, inconsistent in format, and lacks connectivity (Eberendu, 2016). This problem extends to the Lebanese data, posing obstacles for researchers and analysts aiming to analyze the data, derive insights, and compare it with data from other countries. Since data exploration is a challenging task due to disconnection and inconsistency in the format (Idreos et al, 2015), the above challenges make it even more challenging. For example, during covid period, consider the need to visualize covid status in Lebanon by focusing on number of cases, mortality, and population. However, this data is available in two disconnected datasets (one from Ministry of Public Health<sup>1</sup> about deaths and the other from Impact open data<sup>2</sup> about cases) and both do not contain population. In this context, the main challenges are the disconnection, aggregation, and enrichment of the datasets that limit the generation of data-driven insights.

Based on the above challenges, this thesis will invest in answering the following research questions:

- How can we improve data integration and reuse of data in a clean and downloadable format to facilitate data analysis?
- Moreover, how can we contextually expand data on the fly to leverage its value and enhance data exploration?

---

<sup>1</sup> Link: <https://www.moph.gov.lb/en/Pages/8/20380/hospital-based-cause-of-death-statistics>

<sup>2</sup> Link: <https://development.impact.gov.lb/ruraldevelopmentmobile/health/status>

To address these questions, this thesis proposes KGFusionX, a framework to convert distributed and disconnected data into a highly readable, understandable, and interconnected knowledge graph format, with the possibility to enrich and aggregate data through its explicit semantics.

The proposed approach involves the design and implementation of a comprehensive framework inspired by industry-leading frameworks such as LexMa (Tyagi & Jimenez-Ruiz, 2020)), Metaphactory platform (Haase et al, 2019), RML views-based framework (Arenas-Guerrero et al, 2023), and Rhizomer (Garcia et al, 2022). This framework enables the conversion of distributed data into knowledge graphs, harnessing the advantages associated with this powerful data representation technique. The knowledge graph data is stored in a triple store to ease data accessibility and reuse. When transforming data into knowledge graphs, several key benefits are realized. Knowledge graphs provide a representation of data by capturing intricate relationships between entities, thereby facilitating a deeper understanding of the data landscape. Moreover, they enable seamless integration of diverse data sources, ensuring flexible interoperability through explicit data semantics (Johnson et al ,2022).

Furthermore, knowledge graphs will help with the enrichment of the data using linked open data sources. With leveraging contextual understanding and semantic information, knowledge graphs enhance data exploration and discovery, empowering users to uncover hidden patterns and valuable insights. Ultimately, this framework will empower data publishers to organize, analyze, and leverage their data sources more effectively. In addition, the value of knowledge graphs was demonstrated through Streamlit app (Khorasani & Hernández , 2022) specifically tailored for data discovery purposes, enhancing research ability, and boosting data exploration. The interface

enables user to convert the linked data (knowledge graphs) into a structured format (cleaned comma-separated values files) but in a way they can be easily visualized on exploratory data visualization tools such as Tableau or PowerBI platform. The interface serves as a convenient platform for accessing knowledge graph data, providing users with intuitive browsing and selection capabilities.

The interface includes multiple features, appropriate categories are assigned to the data files to aid browsing and searching the datasets. By utilizing categorization users can easily navigate through the repository and locate the specific data they need in order to enhance and enable usability, features such as filtering, sorting, and data. The interface offers users the flexibility to select specific subsets or formats of the data while being kept informed of the download progress. KGFusionX aims to provide smart-self-service access to data. It contributes to empowering users to find and download the information they require conveniently and intuitively (Bodker, 2021). This user-friendly approach will significantly enhance the accessibility and usability of the converted national Lebanese data which include 78 datasets that generated 749,500 Resource Description Framework (RDF) triples from several domains (Health, education, infrastructure, and others), benefiting researchers, analysts, and organizations seeking valuable insights for decision-making processes. Moreover, the generated CSV files can be easily visualized by users using Tableau. KGFusionX was used by data analysts in two use cases to get data and visualize it on Tableau, they leveraged their visuals using the integration and enrichment features. Two successful projects were obtained that focused on the problems of education and tourism in Lebanon. This research provides the following main contributions:

First, it expands the research on knowledge graphs to support data exploration processes, second it offers a novel online data platform that connects previously disconnected datasets that are now publicly available.

### **1.1 Aims and Objectives**

This thesis focuses on facilitating data exploration by enhancing data enrichment and aggregation. Thus, our aims and objectives are to design a framework for:

1. Connecting different datasets from different sources and domains.
2. Implementing a tool that converts distributed data to Knowledge graphs.
3. Facilitating data enrichment using linked open data sources.
4. Integrating different datasets easily and efficiently.
5. Obtaining datasets that can be easily and effectively visualized.
6. Building an interactive interface where users can filter, enrich, integrate, and download data.

### **1.2 Thesis Outline**

Chapter 2 presents a comprehensive literature review of the related work in the field. Thus, this chapter investigates the gap from the previous research and work done about our topic and proposes a possible solution.

Chapter 3 explains the research methodology of the thesis which includes five main steps: Data preparation, RDF generation, RDF repository, CSV Builder, and finally KGFusionX that contains important functionalities such as data enrichment, data integration, and filtering.

Chapter 4 elaborates on the results of the thesis and highlights two main use cases. The use cases highlight the results and how the methodology addressed the problem and the gap in focus.

Chapter 5 concludes the thesis and illustrates the achieved objectives and contributions. Moreover, it draws attention on the limitations and potential future research.

# CHAPTER 2

## LITERATURE REVIEW

### 2.1 Introduction

This part aims to get better understanding of knowledge graphs to explore related work done in this scope. It includes an overview of frameworks for knowledge graphs generation, the use of semantic web approaches in data exploration, and data enrichment approaches using knowledge graphs.

### 2.2 Frameworks for Knowledge graphs generation

Arenas-Guerrero et al. (2023) introduced a novel method for generating knowledge graphs from tabular data using RML views. The authors address the limitations of RML in handling tabular data, particularly in defining complex joins and transformation functions. To overcome these limitations, they propose the use of RML views, which are SQL queries that enable computations on the tabular data before transforming it into RDF. The authors implement this approach in a system called Morph-KGC and conduct an evaluation comparing it to other RML-based systems for knowledge graph generation from tabular data. The evaluation demonstrates that Morph-KGC exhibits superior scalability, enabling the generation of knowledge graphs from larger datasets.

The paper presents three main contributions. First, RML views enhance the transformation capabilities by enabling the definition of complex transformations on tabular data, including joins and aggregations. This results in the generation of more expressive knowledge graphs that effectively capture the relationships between entities in the data. Second, the evaluation of Morph-KGC indicates improved scalability



compared to traditional RML mappings. The efficiency of evaluating RML views allows for the processing of larger datasets, addressing a critical challenge in knowledge graph generation. Finally, RML views offer enhanced maintainability by leveraging SQL, a widely used and understood language. The familiarity with SQL among practitioners simplifies the development and maintenance of RML views, reducing the complexity associated with managing knowledge graph generation processes. In conclusion, this paper presents a promising approach for generating knowledge graphs from tabular data. By utilizing RML views, this method provides improved transformation capabilities, scalability, and maintainability compared to traditional RML mappings. The findings of this paper contribute to advancing the field of knowledge graph generation and offer valuable insights for researchers and practitioners in this area.

Santipantakis et al. (2022) presents a novel approach for generating RDF triples from big data sources. RDF-Gen is a generic mechanism that supports the transformation of data efficiently, even in cases where the velocity of data presents high peaks. It also offers facilities for discovering associations between data from different sources and supporting transformation of modular data sets. RDF-Gen works by means of triple templates, which are similar to SPARQL graph patterns, and contain template variables that are associated with data from the source. Any data conversion, filtering or processing is applied via custom functions, on the template variables. The paper also presents a parallel implementation of RDF-Gen, as well as data transformation workflows that allow variations incorporating RDF-Gen instances, adjusting to the needs of data sources, application areas and performance requirements. RDF-Gen is a valuable tool for anyone who needs to generate RDF triples from big data sources. It is

particularly useful for applications in the areas of data integration, knowledge graphs, and the Semantic Web.

The main features of RDF-Gen are efficiency since it is designed to generate RDF triples efficiently even from high-velocity data sources and scalability due to its ability to handle large datasets and multiple data sources. In addition, it supports a wide range of data formats and ontologies and can be extended with custom functions to support complex data transformation tasks. For that, RDF-Gen is a powerful and flexible tool for generating RDF triples from big data sources. It is an asset for anyone who needs to develop data integration, knowledge graph, or Semantic Web applications.

### **2.3 Frameworks that use Semantic web in data exploration**

Haase et al (2019) provide a comprehensive overview of a platform designed for managing knowledge graphs which caters to a wide range of users, especially data analysts. The paper begins by elucidating the multifaceted challenges involved in knowledge graph management that include but not limited to data integration, efficient storage and querying, and the imperative need for user-friendly visualization. Afterwards, it delves into the architecture of the Metaphactory platform (scalability and adaptability) and explores its key features: data integration capabilities, diverse storage formats, query language support, versatile data visualization tools, and the framework it offers for application development. Also, it provides the readers real-world examples of the platform's use across four application domains that include knowledge graphs which are cultural heritage, the energy industry, and life sciences, are provided. Furthermore, the paper imparts valuable insights garnered from practical experiences in deploying knowledge graph applications within an enterprise context. However, Metaphactory

platform has several limitations especially when dealing with large datasets because the process of setting and managing knowledge graphs is complex and requires expertise in the domain. Moreover, Metaphactory does not offer wide customization options which restricts tailoring it for specific needs and workflows.

In conclusion, the authors discuss the promising future potential of the Metaphactory platform as a valuable tool for managing knowledge graphs in a wide array of application domains.

Tvarožek & Bieliková (2010) highlight in their paper that the Semantic Web is a vision of the World Wide Web in which information is given well-defined meaning. This makes it easier for machines and analysts to process and understand. Also, this paper discusses the challenges of generating exploratory search interfaces for the Semantic Web. It is a promising platform for exploratory search, which is a type of search that involves browsing and exploring data in an open-ended way. The paper begins by reviewing the different types of exploratory search interfaces that have been proposed for the Semantic Web. These interfaces can be classified into three main categories which are keyword-based, view-based, and content-based. Keyword-based interfaces allow users to search for information by entering keywords. View-based interfaces allow users to explore data by browsing through different views or perspectives. Content-based interfaces allow users to explore data by interacting with its underlying content.

After that, the paper then discusses the challenges of generating effective exploratory search interfaces for the Semantic Web which are but not limited to: large volume and complexity of Semantic Web data, the lack of standard data formats and the need to support different user tasks and goals. To overcome these issues, the paper

proposes a framework for generating exploratory search interfaces for the Semantic Web. The framework consists of three main components: a data model, a search engine, and a user interface. The data model represents the Semantic Web data in a way that is easy for the search engine to process. The search engine then uses this data to generate search results. The user interface allows users to interact with the search results and explore the data. Finally, the paper evaluates the framework using a case study of a Semantic Web application for the tourism domain. The results of the evaluation show that the framework is effective in generating exploratory search interfaces for the Semantic Web. The paper concludes by discussing the future of exploratory search on the Semantic Web. The authors believe that the Semantic Web has the potential to make exploratory search more effective and efficient. However, they also acknowledge that there are still several challenges that need to be addressed before this can be achieved such as scalability issues especially for interface generation and end-user exploration.

Nuzzolese et al (2017) presents a novel approach to Linked Data exploration that uses Encyclopedic Knowledge Patterns (EKPs) as relevance criteria for selecting, organizing, and visualizing knowledge. EKPs are discovered by mining the linking structure of Wikipedia and evaluated by means of a user-based study, which shows that they are cognitively sound as models for building entity summarizations. The authors of this paper argue that the traditional approach to Linked Data exploration, which relies on keyword-based search, is not sufficient to address the challenges posed by the heterogeneity and complexity of Linked Data. For that, they propose EKPs as a more effective way to select and organize knowledge, as they capture the semantic relationships between entities in a way that is more intuitive and meaningful to humans. The authors implemented a tool named Aemoo that supports EKP-driven knowledge

exploration. Aemoo integrates data coming from heterogeneous resources, namely static and dynamic knowledge as well as text and Linked Data. It is evaluated by means of controlled, task-driven user experiments, which show that it is able to provide relevant and serendipitous information to users. The paper makes a number of contributions to the field of Linked Data exploration. It introduces the concept of EKPs as a new way to represent and reason about knowledge in Linked Data. Also, it proposes a novel approach to Linked Data exploration that is based on EKPs. Finally, it implements a tool, Aemoo, that supports EKP-driven knowledge exploration. The evaluation results are promising and suggest that Aemoo is a promising tool for Linked Data exploration. However there are some of the limitations of the paper such as the user study was conducted with a small number of participants, so the results may not be generalizable to a wider population, the evaluation of Aemoo was conducted in a controlled setting, so it is not clear how it would perform in a real-world setting, and the paper does not discuss how EKPs can be used to support other tasks, such as question answering and recommendation. The evaluation results are promising and suggest that Aemoo is a promising tool for Linked Data exploration.

Thalhammer, Lasierra & Rettinger (2016) propose a link-based approach for the relevance-oriented summarization of knowledge graph entities. The approach, called LinkSUM, optimizes the combination of the PageRank algorithm with an adaptation of the Backlink method together with new approaches for predicate selection. The paper conducted both quantitative and qualitative evaluations of LinkSUM. The quantitative evaluation showed that LinkSUM significantly outperforms the state-of-the-art entity summarization approach in terms of both relevance and coverage. The qualitative evaluation showed that LinkSUM summaries are more concise and easier to understand

than the summaries generated by the state-of-the-art approach. The paper concludes that LinkSUM is a promising approach for the summarization of knowledge graph entities. It is lightweight and efficient, and it can be easily adapted to different knowledge graphs. The main contributions of this paper are proposing a novel link-based approach for entity summarization, conducting comprehensive evaluations of the proposed approach to show that it outperforms the state-of-the-art and opening up new possibilities for the summarization of knowledge graph entities.

García, López-Gil, & Gil (2022) shed light on a new web application: “Rhizomer”. Rhizomer is an open-source web application for exploring knowledge graphs that is designed to be easy to use, even for users with no prior knowledge of knowledge graphs or semantic data and query languages. It provides three main features which are overview, zoom and filter, and details-on-demand. Overview allows users to get a high-level view of the knowledge graph, including the main classes and relationships between them. With Zoom and filter feature, Users can zoom in on specific parts of the knowledge graph and filter out irrelevant information. Details-on-demand feature help Users get more detailed information about any node or relationship in the knowledge graph. Rhizomer has been used in a variety of scenarios, including research, commercial projects, and teaching. It is a valuable tool for anyone who needs to explore and understand knowledge graphs.

García, López-Gil & Gil (2023) describe in their paper how the authors used the Benchmark for End-User Structured Data User Interfaces (BESDUI) to guide the development of a new semantic knowledge graph exploration tool called RhizomerEye. The authors first evaluated the predecessor of RhizomerEye using the BESDUI benchmark. They then used the results of the evaluation to identify areas where the tool

could be improved. They then made the necessary changes to the tool and evaluated it again using the BESDUI benchmark.

The results of the second evaluation showed that RhizomerEye had improved significantly in terms of user experience. The authors also conducted a user evaluation of RhizomerEye using the same dataset and tasks as the BESDUI benchmark. The results of the user evaluation were comparable to those of the BESDUI benchmark, especially for users with knowledge about semantic technologies. The authors conclude that the BESDUI benchmark can be used effectively to guide the development of semantic knowledge graph exploration tools. They also note that it is important to evaluate the user experience of these tools with real users, in addition to using benchmarks.

Vogt (2023) argues that the FAIR principles, which are a set of guidelines for making data and metadata findable, accessible, interoperable, and reusable, do not adequately address the human-actionability of data and metadata. Vogt suggests adding a fifth principle, "Human Explorability," to the FAIR principles to make them Fairer. He also discusses the importance of cognitive interoperability, which is the ability of humans to understand and interact with data and metadata in a meaningful way. Vogt argues that cognitive interoperability is essential for making data and metadata truly FAIR and Fairer.

In the second part of the paper, Vogt introduces semantic units, which are a way of structuring knowledge graphs into identifiable and semantically meaningful subgraphs. He argues that semantic units can increase the human explorability and cognitive interoperability of knowledge graphs. Vogt concludes the paper by discussing how semantic units can be used to develop innovative user interfaces that support exploring

and accessing information in knowledge graphs by reducing its complexity to what currently interests the user.

Al-Tawil, Dimitrova & Thakker (2020) propose a novel approach to facilitate users' exploration of data graphs in a way that leads to expanding the user's domain knowledge. The approach is based on the subsumption theory of meaningful learning, which postulates that new knowledge is grasped by starting from familiar concepts in the graph which serve as knowledge anchors from where links to new knowledge are made. The authors introduce a new framework for generating exploration paths for knowledge expansion using knowledge anchors. The framework consists of two main components: A knowledge anchor detection algorithm (KADG) for automatically identifying knowledge anchors in a data graph and a subsumption algorithm that utilizes KADG to generate exploration paths for knowledge expansion.

The framework is evaluated in a task-driven experimental user study in the music domain. The results show that exploration paths using knowledge anchors and subsumption lead to significantly increase the users' conceptual knowledge and better usability compared to free exploration of data graphs. The paper makes a significant contribution to the field of data graph exploration by proposing a novel approach that is grounded in educational theories and empirically shown to be effective in facilitating users' knowledge expansion.

Ermilov, et al. (2017) presents GENESIS, a generic RDF data access interface that can be deployed on top of any knowledge base and search engine with minimal effort. GENESIS allows for the representation of RDF data in a layperson-friendly way, which is facilitated by its modular architecture for reusable components. These components include a generic search back-end, together with corresponding interactive user



interface components based on a service for similar and related entities as well as verbalization services to bridge between RDF and natural language.

GENESIS has a few potential benefits, as it can make RDF data more accessible to a wider range of users, including those who are not familiar with RDF or SPARQL. Also, it can help users to better understand the results of RDF queries by presenting them in a more layperson-friendly way. Moreover, it can facilitate the navigation and exploration of RDF data. The authors of the paper evaluate GENESIS using several different datasets and show that it can be used to effectively implement a variety of RDF data access tasks. Overall, GENESIS is a promising new approach to RDF data access that has the potential to make RDF data more accessible and useful to a wider range of users.

Scarpato & Alessio (2017) talks about SAGG which is a knowledge-based visualization system that exploits user-submitted examples to automate the visualization process. It consists of three main components which are SAGG VIPS algorithm, SAGG Semantic Search Algorithm, and SAGG Structuring & Formatting Fresnel-based approach which is a configuration file composer module able to create groups, lenses, and formats according to the Fresnel standard language.

The SAGG system works by first analyzing the user-submitted examples to identify the key entities and relationships in the data. It then uses this information to generate a semi-automatic GUI that allows the user to visualize the data in a variety of ways. SAGG has been evaluated on several real-world datasets, and the results show that it can be used to effectively visualize complex linked data in a way that is both informative and easy to understand. One of the key advantages of SAGG is that it does not require users to have any prior knowledge of linked data visualization. This makes it

a valuable tool for a wide range of users, including domain experts, researchers, and journalists. SAGG is still under development, but it has the potential to revolutionize the way that linked data is visualized and explored.

In their paper, Ikkala et al. (2022) present SAMPO-UI, a software framework designed for developing user interfaces in semantic portals. SAMPO-UI enables end-users to explore Linked Data knowledge graphs from multiple perspectives using a two-step usage cycle of faceted search and data analysis tools. For software developers, SAMPO-UI offers a customizable and user-friendly environment, leveraging state-of-the-art JavaScript libraries and data from SPARQL endpoints to reduce coding effort. SAMPO-UI has been successfully utilized in several projects, particularly in the Cultural Heritage domain, resulting in the creation of numerous portals with tens of thousands of end-users. The framework is open-source, available on GitHub under the MIT License, and its contributions include enhanced exploration capabilities and efficient interface development for semantic portals.

SAMPO-UI allows users to explore Linked Data knowledge graphs quickly and easily. The framework also makes use of state-of-the-art JavaScript libraries, which ensures that it is responsive and user-friendly. Additionally, SAMPO-UI can reduce coding effort by leveraging data from SPARQL endpoints. However, SAMPO-UI is still under development, and there are some limitations to the framework. For example, the two-step usage cycle can be somewhat restrictive, and the framework does not support all types of Linked Data knowledge graphs.

McCusker & McGuinness (2023) introduce Whyis 2, a new open-source framework specifically designed for knowledge graph development and research. The authors aim to create an accessible and extendable platform by incorporating various

features that cater to the needs of KG development and research. Whyis 2 offers a modular architecture that allows for easy customization and interchangeability of different components. Additionally, it provides a robust toolset for KG development, encompassing tasks such as data loading, cleaning, and validation. Furthermore, the framework includes advanced features specifically tailored for KG research, including entity alignment and link prediction capabilities. To illustrate the practical application of Whyis 2, the paper presents several case studies demonstrating its effectiveness in building KGs from scratch, extending existing KGs, and conducting KG research. In conclusion, the authors discuss the advantages of utilizing Whyis 2 for KG development and research. The open-source nature of the framework makes it accessible to a wide range of users, while its modular architecture enables effortless customization and extension. Moreover, the rich set of tools and features provided by Whyis 2 renders it suitable for both KG development and research. Overall, the paper "Whyis 2: An Open-Source Framework for Knowledge Graph Development and Research" serves as a valuable resource for researchers and developers interested in constructing and utilizing KGs, offering a detailed overview of Whyis 2 alongside compelling case studies that showcase its potential in KG development and research endeavors.

Escobar et al. (2020) focuses on enhancing the accessibility and analysis of statistical data from public repositories through the application of Semantic Web standards, specifically RDF and SPARQL. The foundation of this framework lies in the utilization of the RDF Data Cube vocabulary, a recognized standard for representing multidimensional data within RDF structures. The proposed framework comprises distinct phases, commencing with a meticulous evaluation of data quality grounded in criteria aligned with the RDF Data Cube vocabulary. Subsequently, data enrichment

ensues through interlinking with external repositories like Wikidata and GeoNames. The enriched data is then published in RDF format, employing the RDF Data Cube vocabulary, followed by its strategic exploitation. This exploitation involves the provision of user-friendly dashboards tailored for non-expert users and the establishment of a SPARQL endpoint catering to expert users.

Validation of the framework is achieved through a case study centered on the Barcelona Open Data platform, demonstrating its efficacy in augmenting Linked Open Data by simplifying the analysis of multidimensional models. The benefits derived from adopting the multidimensional model approach based on the RDF Data Cube vocabulary are manifold. It enhances the comprehension and analysis of intricate datasets, facilitates the creation of aggregated perspectives that unveil trends and patterns, enables cross-source data comparison, and fosters the seamless reuse and sharing of data. In conclusion, this paper stands as a noteworthy contribution to the realm of Linked Open Data, proposing a robust framework for the publication and exploitation of multidimensional data using Semantic Web standards, thereby advancing the accessibility and usability of such data within the research community.

## **2.4 Data Enrichment using Knowledge graphs**

Furthermore, Quattrini, Pierdicca & Morbidoni (2017) presents an intriguing perspective on leveraging the semantic web to enrich Hyper-Building Information Modeling (HBIM) data. It introduces a novel approach to enhancing data quality and interoperability within HBIM. It advocates the benefits of employing the semantic web to offer a more structured framework for representing and sharing HBIM data. A standout feature is their proposed method for automatically enriching HBIM data with

semantic web knowledge, showcased through rigorous experiments demonstrating a significant enhancement in HBIM data quality. Key points highlighted in the paper include the progressive role of HBIM reliant on knowledge graphs, the structured approach of the semantic web for HBIM data representation, and the innovative automated enrichment method for HBIM data. Ultimately, this paper makes a substantial contribution to the HBIM field, shedding light on potential impacts in data representation and its practical applications.

## **2.5 Gap and proposed Solution**

Overall, it is clear that most of the previous work focus on generating knowledge graphs from distributed data using several frameworks (Liu et al., 2022) or knowledge graph exploration and search (Scarpato & Alessio ,2017 ; Haase et al, 2019; Al-Tawil, Dimitrova & Thakker, 2020). However, there is still few gaps that need to be tackled, first most of the frameworks including the above ones focus on knowledge graphs generation and exploration and only few commit to data enrichment, however using knowledge graphs to enrich data is still limited to specific fields such as HBIM data (Quattrini, Pierdicca & Morbidoni, 2017) and they do not take into account distributed data (tabular). In other words, there is no framework that covers KG generation, data integration and data expansion at the same time. Thus, we propose focusing on facilitating data enrichment and integration using knowledge graphs which will not only help users in these two objectives, but also, they will be able to visualize the enriched and integrated data on Tableau.

# CHAPTER 3

## RESEARCH METHODOLOGY

In this research, we propose a comprehensive framework to facilitate data integration and enrichment using knowledge graphs. Figure 1 Represents the high-level research methodology framework. For data integration, our first step is data preparation because we needed to have data in our hands to evaluate if our framework will yield efficient results or not. It includes data collection and then data wrangling to be able to process it. Next, we convert the data collected to Resource Description Framework (knowledge graphs) following the linked data cube ontology, then we store the data in a repository (triple store). Then, the RDF data is extracted based on data domain and category where they are presented on the tool “KGFusionX”, and using the connections generated by the knowledge graphs (URIs and links) the user is able to integrate any two or more data frames using Full Outer Joint technique. The interface offers the users the ability to filter, enrich through contextual enrichment from linked open data sources (e.g. Dbpedia) and download the merged datasets.

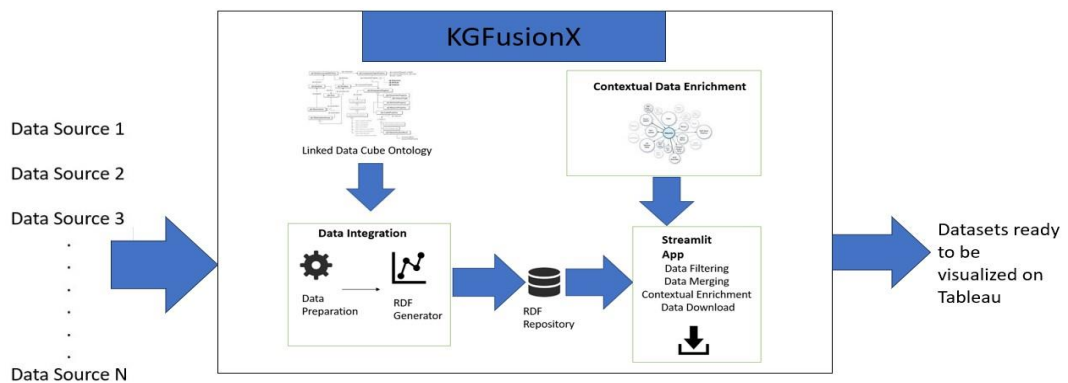


Figure 1 Research Methodology Framework (high level)

### 3.1 Data Sources

One of the main challenges in developing countries is the slow adoption of digitalizing of services at the government level. For example, the Lebanese government and all its related entities (ministries, municipalities, etc.) are still far from digitalizing their services. We focused in this thesis on the available data on the websites of the ministries that publish their data in a non-organized manner. The data found in different formats (HTML, csv, pdf, excel, etc.). Not only that, but also the websites are not accessible for several days. We have proceeded in working on the available data on the websites of ministries and searched for other data sources such as Central Administration of Statistics (CAS)<sup>3</sup>, Impact open data<sup>4</sup>, and Brite<sup>5</sup>. Of course, in addition to other data sources that are more international such as the United Nations (World Health Organization)<sup>6</sup>.

The initiation was from the ministry of public health, where I have found "Surveillance data". This data provides users with the reported cases of mandatory notifiable diseases.

The data is displayed in the national and governorate levels for the past 17 years. This distribution of data makes it hard to explore later, because about notifiable diseases only we have more than 150 HTML pages making them hard to analyze! This also sheds light on the problem of data disconnection.

---

<sup>3</sup> Link to "CAS" website: <https://www.cas.gov.lb/>

<sup>4</sup> Link to "Impact Open Data" website: [https://impact.cib.gov.lb/home#open\\_data\\_section](https://impact.cib.gov.lb/home#open_data_section)

<sup>5</sup> Link to "Brite" website: <https://brite.blominvestbank.com/>

<sup>6</sup> Link to "WHO" website: <https://data.who.int/countries/422>

Communicable Diseases Surveillance

Table Beirut - A

Reported cases by time - Year 2021

as on 09/03/2022

	TOTAL	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
<b>Vaccine Preventable Diseases</b>													
Acute Flaccid Paralysis	2	0	0	0	1	0	1	0	0	0	0	0	0
Acute Poliomyelitis	0	0	0	0	0	0	0	0	0	0	0	0	0
Diphtheria	0	0	0	0	0	0	0	0	0	0	0	0	0
Measles	3	0	0	0	0	1	0	0	2	0	0	0	0
Mumps	0	0	0	0	0	0	0	0	0	0	0	0	0
Pertussis	0	0	0	0	0	0	0	0	0	0	0	0	0
Rabies	0	0	0	0	0	0	0	0	0	0	0	0	0
Rubella	0	0	0	0	0	0	0	0	0	0	0	0	0
Tetanus	0	0	0	0	0	0	0	0	0	0	0	0	0
Tetanus neonatal	0	0	0	0	0	0	0	0	0	0	0	0	0
Viral Hepatitis B	1	0	1	0	0	0	0	0	0	0	0	0	0
<b>Food &amp; Water Borne Diseases</b>													
Breastfeeding	10	2	1	1	0	0	0	0	0	1	1	3	1
Cholera	0	0	0	0	0	0	0	0	0	0	0	0	0
Dysentery	120	5	2	3	5	9	7	20	16	27	9	13	4
Food Poisoning	4	0	0	1	0	0	0	2	1	0	0	0	0
Hydratic Cyst	0	0	0	0	0	0	0	0	0	0	0	0	0
Parasitic Worms	0	0	0	0	0	0	0	0	0	0	0	0	0
Typhoid	0	0	0	0	0	0	0	0	0	0	0	0	0
Typhoid Fever	0	0	0	0	0	0	0	0	0	0	0	0	0
Viral Hepatitis A	7	1	2	1	0	0	0	1	0	1	1	0	0
<b>Other Diseases</b>													
Bilharziasis	0	0	0	0	0	0	0	0	0	0	0	0	0
Creutzfeldt Jakob Disease	0	0	0	0	0	0	0	0	0	0	0	0	0
Ebola	0	0	0	0	0	0	0	0	0	0	0	0	0
Goorrhoea	1	0	0	1	0	0	0	0	0	0	0	0	0

Figure 2 An HTML page

After that, I proceeded in data collection from the ministry of public health, and I collected datasets related to mortality such as hospital deaths and Covid deaths. They contain number of death cases distributed among Age-groups and time (monthly data). These datasets are in PDF format, this puts another challenge of dealing with different files format which also makes it hard to explore and analyze. Moreover, I collected data about Cancer which is also in PDF format distributed also on Age-groups and cancer type.

In the attempt to collect data from different sources, and different sectors (not only the health sector), I have also explored several other websites, however I have found good datasets on the Impact Open Data website. The IMPACT Open Data website is a project by the Central Inspection aimed at providing public access to the extensive data collected from a collaborative, nationwide online data collection effort involving various government ministries and local municipalities. IMPACT stands for the Inter-Ministerial and Municipal Platform for Assessment, Coordination, and Monitoring. It contains data related to health, education, social, and economic sectors distributed by areas in terms of towns, and districts. These datasets are in excel format, and some of its columns are in Arabic language such as districts and towns columns.



This adds a new challenge to data analysts in exploring these datasets which is the multilingual challenge.

Furthermore, data has been collected from other sources such as brite, world health bank, and world health organization. All these datasets make our data collection diverse as much as possible. In total, we have collected 78 datasets, from several data sources and different domains (health, education, social, infrastructure, economic, and other).

During this process, several challenges have occurred. For instance, the diseases data for 2020 is not available, also when I started the data for 2022 (diseases) was available, however it disappeared before me to save it in the triple store, since I was still practicing coding on it. For that, and in the sake of preserving the data, I downloaded all the data that could be downloaded and used. And for the html pages, I have saved them in an online repository on GitHub, in addition to converting them to csv files and saving the csv files. Also, some datasets such as hospitals deaths are not published after 2021. Regarding other datasets, such as impact datasets, they miss the “time” thus we do not know the date of this data and if it is updated regularly or not.

### **3.2 Data Preparation**

Data cleaning is an important step in data preparation (Mishra et. al, 2020), and similar to any other raw data, it has missing values, unclear format, and some redundant values, for that it is a must for us to preprocess the data. The first step is to unify the format of the datasets because there are different datasets from different sources that could be analyzed with the diseases datasets that are HTML such as data related to mortality that are available in pdf format, and health resources datasets that are

available in excel format. For that, the choice is to make them all csv files because they are easily converted to RDF, and it is the most suitable format for python to work with. Also, in case the website where the HTML file or other files are located is down for any reason, we secured the data that we have collected since we placed our HTML files in an online repository, and other datasets in a drive.

### ***3.2.1 HTML TO CSV***

The next step involved preprocessing with the disease's datasets, these datasets are in HTML format; thus, they are hard to process and cleaned, in addition to that they are disconnected. In processing the HTML files, I have used BeautifulSoup library in python (Patel & Patel, 2020). This library parses html files and is able to convert them to csv files efficiently, we have used one code for all the html files, here is an example:

```
# Parse the HTML content with BeautifulSoup
soup = BeautifulSoup(html_content, 'html.parser')

# Find the table you want to extract data from
table = soup.find('table')

# Create a CSV file and write the header row
with open('output.csv', mode='w', newline='') as csv_file:
    writer = csv.writer(csv_file)
    header_row = [header.text for header in
table.find_all('th')]
    writer.writerow(header_row)

# Write each row of data to the CSV file
for row in table.find_all('tr'):
```

```

        data_row = [data.text for data in
row.find_all('td')]
        writer.writerow(data_row)
        import pandas as pd

# Read the CSV file into a DataFrame
df = pd.read_csv('output.csv')

```

### 3.2.2 PDF TO CSV

After that, for the datasets that are in PDF format such as the mortality datasets, we used Tabula library in python (Rosen, 2019) , it parses the pdf documents and is able to convert them to csv files, similar to what is done with HTMLs, we developed one code to parse pdf files:

```

import tabula

# Specify the path to your PDF file
pdf_path = "/content/HMS Lebanese Mortality 2018 by Age group
and Month.pdf"

# Specify the page number(s) containing the table(s) you want
to extract
# You can also use 'all' to extract tables from all pages
pages = 'all'

# Extract tables from the PDF and save them as a list of
DataFrame objects
tables = tabula.read_pdf(pdf_path, pages=pages)

# Iterate over the extracted tables and save them as CSV files
for i, table in enumerate(tables, start=1):

```

```
table.to_csv(f"output_table_{i}_2018.csv", index=False)

print("PDF to CSV conversion completed successfully!")
```

After unifying the format of our datasets, it is essential to clean them. We have removed empty columns and rows that have been formed due to the conversion, also in case of any duplication, we have removed them.

Regarding null values, is dependent on which dataset and which column they are present. For example, if the null values are in the death cases or disease cases, we assume that we have no cases thus it is zero. However, for other columns, we treated them as missing values and they are filled case by case (by mode or mean). But I have tried my best to preserve the datasets because our main goal in this thesis is to facilitate data exploration.

### **3.3 RDF Generator: CSV to RDF**

In order to help users benefit from the available data in a more efficient way, and for them to access the data more easily, and in an attempt to find links between different datasets to connect them, and after doing research, I found that the best way to do so is knowledge graphs (Blomqvist et. al, 2021) as they connect data with each other using linked data ontologies as triples: Subject, Object, Predicate (Chen et. al, 2020).

But we need to maintain the links between different data sources and international data by having an open linked data source that could help us. The choice was Dbpedia as it is an open source linked data that has huge amount of data since it is directly linked with Wikipedia (Lehmann et al, 2015). Thus, we have sent in our code calls to Dbpedia Lookup to retrieve the URIs of areas, and entities found in our data that could be linked

together. Making such code was not easy, this is because of the typing errors for some entities which makes the code retrieve a totally different unique resource identifier (Yulianti et al,2021).

Also, sometimes the same word has more than one meaning which leads to having two URIs that are very similar to each other and thus our code may retrieve the wrong URI. For example, in some datasets there is a disease named “Mumps”, on Dbpedia there is a URI for this disease and there is a URI for a programming language named “MUMPS”. This is called word sense disambiguation which is a common challenge.

To illustrate more, it is better to give an example. In some datasets, the data is grouped as governorates and in other as districts, and in others at the national level. So, as we know, districts are part of governorates and governorates are part of the country (e.g. Lebanon). Here comes the importance of having the Dbpedia links of each region so that we can group different datasets in terms of location at a later stage. Not only that, but also if the user is interested in knowing more about a certain region, he can follow the link and he will find a full description and other data on the Dbpedia link that we have provided. Another issue that any data analyst may face is the multilingual issue. In some cases, Beirut is typed in English, however in other cases it is typed in Arabic (such as the IMPACT datasets). Dbpedia can potentially solve this issue because we are able to retrieve using the Arabic values through the same Dbpedia URI retrieved from the English value and help with in unifying the data (Bouziane et al, 2020).

Consequently, Dbpedia URIs serve as the connectors between our data and external data. Also, it solves an important challenge which is “typing in different ways.” For instance, we have two different datasets that talk about Covid, in the first one, it is

written corona virus, while in the second covid-19. Another example would be about diseases datasets where the version of 2023 is based on Districts, while that of 2021 is based on Governorates. And some districts have the same name as their governorates such as Nabatiyeh Governorate and Nabatiyeh District and there also exist Nabatiyeh City. This makes it challenging for the data and here comes the power of knowledge graphs to differentiate between them (Fafalios & Tzitzikas, 2019). For that reason, when we are sending a call to Dbpedia, we are specifying the entity type we are looking for. For example, for Districts we are specifying in our code to retrieve only the URIs that are districts: `“search?/typeName=district&query”`

Now our csv files are ready to become knowledge graphs, but we must use an ontology to convert our data to RDF and to maximize data reuse (Wilkinson et al, 2016). After going over literature, I have found that the best ontology to follow is the statistical core vocabulary (SCOVO) as it fits our data (part of our data is statistical) and is easy to use. However, after trying to go over its specific words, it turned out that it's no longer available as its website is out of order: <http://vocab.deri.ie/scovo>). Thus, I started searching for alternatives and the option was “The RDF Data Cube Vocabulary” that is because it is powerful and useful in publishing multi-dimensional data, such as statistics, so that we would be able to link our data to related data sets and concepts (Tennison et al, 2012). Figure 3 shows the ontology schema we have used in building the knowledge graphs. We extended the linked data cube ontology which is the “*Dataset Category*” that is part of the “*Data Domain*” (highlighted in red) in order to connect datasets that come from different websites and sources but have the same domain. For example, a dataset about cholera disease from the website of Ministry of Public Health belongs to Health domain. Also, a dataset from World Health

Organization about mortality belongs to health domain. Thus this extended ontology connects these datasets together.

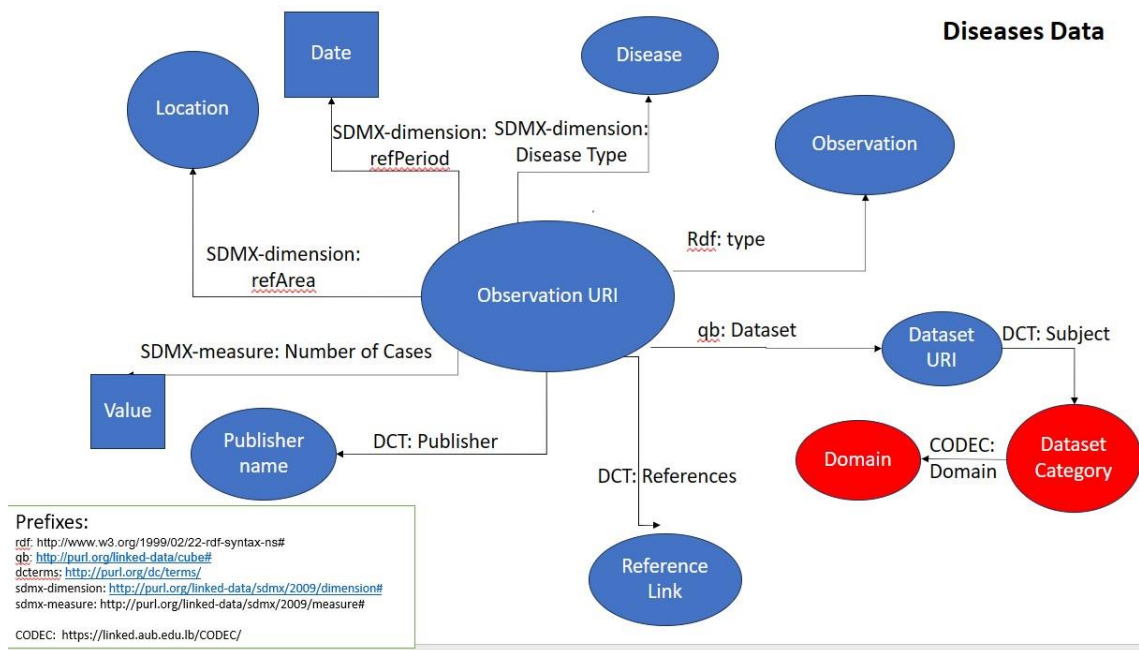


Figure 3 Ontology Schema

After going over the ontology , I started the conversion of csv files to RDF. We converted the distributed data we have collected into knowledge graphs using python. This process involves some challenges especially when we are dealing with different datasets that contain different structures. It was necessary to do a generic code that can fit many datasets in order to decrease the time needed to do the whole process. This code can handle different datasets of different number of columns with different types (integers, strings, float, dates, etc..) (Found in appendix).

After doing the conversion, it was a must to improve the form of the generated knowledge graphs so that they become tidy and readable and easy to be understood by any user. Also, this process was refined several times because we have worked as much

as we can on improving our RDFs and making them rich and valuable and aligned with the ontology. Here is a sample RDF for an Observation:

```
<http://linked.aub.edu.lb/CODEC/observation/Beirut-2012-07-Plague>
  a                               qb:Observation ;
  <http://purl.org/dc/terms#publisher>
    "Ministry of Public Health" ;
  <http://purl.org/dc/terms#references>

  "https://www.moph.gov.lb/userfiles/files/Esu_data/Esu_pastyears/Beirut2012.htm" ;
  qb:dataset
  <http://linked.aub.edu.lb/CODEC/Dataset/Beirut-2012-07-Plague> ;
  sdmx-dimension:disease
  <http://dbpedia.org/resource/Bubonic_plague> ;
  sdmx-dimension:refArea
  <https://dbpedia.org/page/Beirut> ;
  sdmx-dimension:refPeriod      <http://server/unset-base/07-2012> ;
  sdmx-measure:Number_of_cases "0" .
```

To help understand how an observation looks like, figure 4 below shows an observation as a diagram.

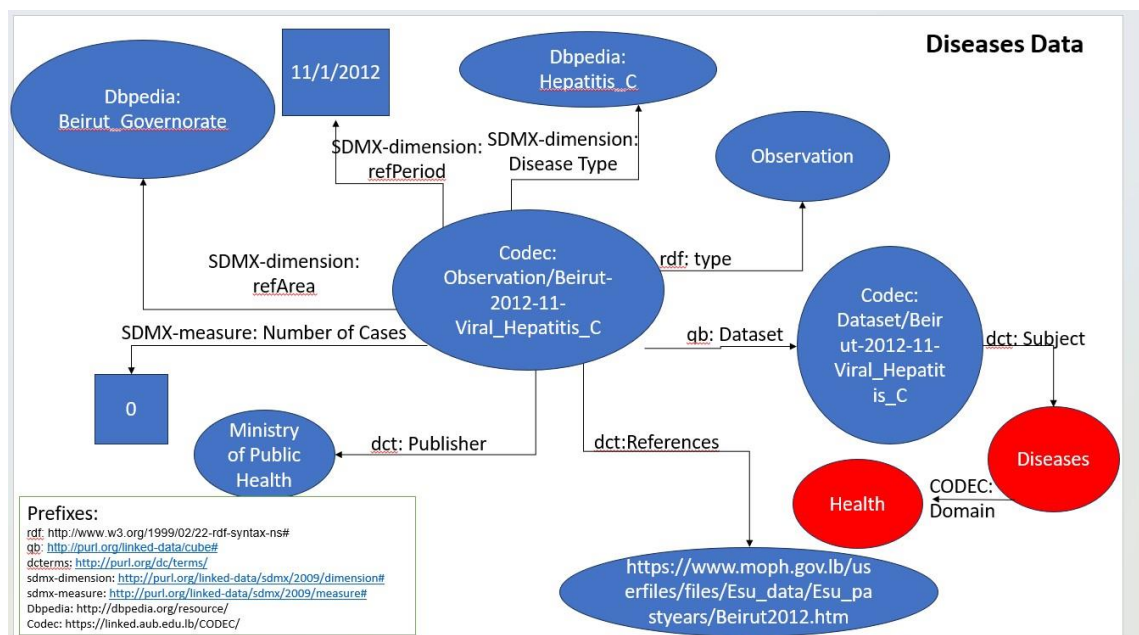


Figure 4 Sample RDF Observation



In total, 749500 triples were generated as shown in table 1.

Table 1 Representing the datasets used, their sources, and number of triples generated

<b>Data Domain</b>	<b>Number of Datasets</b>	<b>Data Sources</b>	<b>Number of RDF triples generated</b>
Health sector	49	MOPH, Impact, Brite, World Health organization	530125
Education sector	4	Impact, MOE	24060
Social sector	8	CAS, Open Data Lebanon, Impact	66215
Economic sector	6	MOF, Impact	40588
Infrastructure	10	Impact	81930
Other	1	Impact	6132
<b>Total</b>	<b>78</b>		<b>749050</b>

### **3.4 Storing RDFs in a Triple Store**

RDF data are usually stored in a Triple Store as it is the most efficient way to store them (Moresy et al, 2012). In this project, we are going to use Fuseki endpoint as a triple store to store our RDF data since it is efficient. Fuseki has several advantages such as its convenient RDF storage space and high performance (Tianyi et al, 2020).

Storing RDF data in the triple store has shown us a new challenge. It is due to some Arabic characters in the RDF data, because storing them in the triple store makes them encoded and thus unreadable by the users. To overcome this challenge, we had a work around by matching Arabic values with their corresponding English values in some datasets such as impact datasets, because the impact website there is a list of some columns such as Town column in both languages, thus we used those common values to do the matching and replacement.

We have used SPARQL Wrapper in order to upload the RDF data to the triple store as it is an easy, straightforward way:

```

from SPARQLWrapper import SPARQLWrapper, RDF
import ssl

ssl._create_default_https_context =
ssl._create_unverified_context

# Set up SPARQL endpoint and return format
sparql =
SPARQLWrapper("https://linked.aub.edu.lb:8080/fuseki/codec_test2
")
sparql.setReturnFormat(RDF)

# Load the contents of the RDF file
with open('/content/Agriculture.rdf', 'r', encoding='utf-
8') as file:
    rdf_content = file.read()

# Set the SPARQL query with the RDF data
sparql.setQuery(f"""
    PREFIX qb: <http://purl.org/linked-data/cube#>
    PREFIX sdmx-dimension: <http://purl.org/linked-
data/sdmx/2009/dimension#>
    PREFIX sdmx-measure: <http://purl.org/linked-
data/sdmx/2009/measure#>
    PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
    PREFIX dct: <http://purl.org/dc/terms#>
    INSERT DATA {{
        {rdf_content}
    }}
""")

# Execute the SPARQL query

```

```
results = sparql.query().convert()

# Process the results as needed
print(results)
```

### 3.5 Presenting the data to the user: “KGFusionX.”

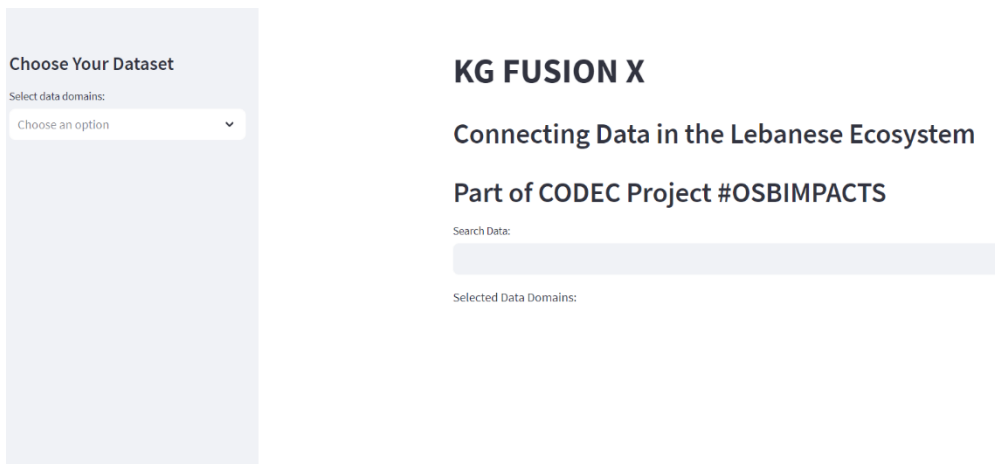


Figure 5 “KGFusionX Interface”

We embark on the crucial task of crafting an intuitive and engaging user interface (UI) that seamlessly connects our users with the wealth of data at their fingertips. This UI is not just a design; it's a gateway to the information they seek, offering an immersive and user-centric experience. This is because this interface uses the power of knowledge graphs in presenting the data.

Using Streamlit, we have developed and designed our interface since it is easy to work with, especially since we use python language in its development. The new data frames presented on Streamlit are created using the power of knowledge graphs (Jares & Klimek, 2021). So, in what ways our csv files are better than the original format?

1. First of all, all the data is now available in one format which is “csv” instead of having multiple formats (HTMLs, CSV, Excel, PDFs, etc..)
2. Second, the data is now created based on the user’s preferences in terms of choosing the main data domain, and the columns he wants, in addition to the filtering criteria.
3. Third, Data frames can be integrated using the full outer joint technique by the user due to the connections generated by knowledge graphs.
4. Fourth, the user is able to enrich the data frames with data that can be retrieved from Dbpedia using sparql queries.
5. Fifth, our generated csv files can be visualized on visualization tools such as tableau thus we have enhanced the reusability of our data in addition to the exploratorily of the data.

Let us delve into the key elements and features that define our user interface:

- **Search Feature:** Now, if users are looking for data related to specific thing (covid-19) for example, but there is no dataset under this name (or typed in a different way) they can benefit from the “search” feature where they can type any word they want and they will get the names of datasets that contain this word and the number of occurrences.
- **Datasets are distributed among domains:** Our interface are unique due to several reasons, first of all, the users are able to choose the data domain(s) they are interested in. After that, they will get the categories available in the domains they have chosen, finally the datasets in these categories will be available for them to choose from.

- Data Filtering:** Now, after the users have chosen the datasets they want, they are now able to filter the columns that contain the dimensions of the dataset. Then, they could choose the dimensions they are interested in to filter. These dimensions include but are not limited to area (district, governorate, country), period, age group, disease, etc. Moreover, there is an option for the users to exclude the columns they are not interested in having in the dataset. Finally, and since our data contains Dbpedia URIs in addition to Observation and Dataset URIs, we provided an option for the users to convert these URIs to Labels as this will facilitate the visualization at a later stage.

**Connecting Data in the Lebanese Ecosystem**

Part of CODEC Project #OSBIMPACTS

Search Data:

Add more data about diseases to Diseases\_Mumps

Choose an option

Selected Data Domains: Health Data

Selected Datasets for Diseases: Mumps

Filtered Data in Diseases/Mumps:

	↑ <a href="http://purl.org/linked-data/cube#dataset">http://purl.org/linked-data/cube#dataset</a>	<a href="http://purl.org/linked-data/sdmx/2009/din">http://purl.org/linked-data/sdmx/2009/din</a>
12	<a href="http://linked.aub.edu.lb/CODEC/Dataset/Beirut-2006-01-Mumps">http://linked.aub.edu.lb/CODEC/Dataset/Beirut-2006-01-Mumps</a>	<a href="https://dbpedia.org/page/Beirut">https://dbpedia.org/page/Beirut</a>
13	<a href="http://linked.aub.edu.lb/CODEC/Dataset/Beirut-2006-02-Mumps">http://linked.aub.edu.lb/CODEC/Dataset/Beirut-2006-02-Mumps</a>	<a href="https://dbpedia.org/page/Beirut">https://dbpedia.org/page/Beirut</a>
14	<a href="http://linked.aub.edu.lb/CODEC/Dataset/Beirut-2006-03-Mumps">http://linked.aub.edu.lb/CODEC/Dataset/Beirut-2006-03-Mumps</a>	<a href="https://dbpedia.org/page/Beirut">https://dbpedia.org/page/Beirut</a>
15	<a href="http://linked.aub.edu.lb/CODEC/Dataset/Beirut-2006-04-Mumps">http://linked.aub.edu.lb/CODEC/Dataset/Beirut-2006-04-Mumps</a>	<a href="https://dbpedia.org/page/Beirut">https://dbpedia.org/page/Beirut</a>
16	<a href="http://linked.aub.edu.lb/CODEC/Dataset/Beirut-2006-05-Mumps">http://linked.aub.edu.lb/CODEC/Dataset/Beirut-2006-05-Mumps</a>	<a href="https://dbpedia.org/page/Beirut">https://dbpedia.org/page/Beirut</a>

Figure 6 Choosing your dataset and filtering functionalities

### 3.6 Data Integration

The users will also be able to merge datasets together using “Full Outer Join” technique which was made possible by the URIs as keys for connecting different tables and datasets. Although merging datasets in one csv file might generate unclean dataset, it would help the users (especially the ones who are expert in some areas) to have the data they want in one place. Also, they can clean the merged data if they find this necessary.

Merged Data:

	on#disease	http://purl.org/linked-data/cube#dataset	http://purl
1,039		http://linked.aub.edu.lb/CODEC/Dataset/North-2021-08-Typhus	Ministry of
1,040		http://linked.aub.edu.lb/CODEC/Dataset/North-2021-09-Typhus	Ministry of
1,041		http://linked.aub.edu.lb/CODEC/Dataset/North-2021-10-Typhus	Ministry of
1,042		http://linked.aub.edu.lb/CODEC/Dataset/North-2021-11-Typhus	Ministry of
1,043		http://linked.aub.edu.lb/CODEC/Dataset/North-2021-12-Typhus	Ministry of
1,044		http://linked.aub.edu.lb/CODEC/Dataset/Beirut-2007-01-Viral_Hepatitis_C	Ministry of
1,045		http://linked.aub.edu.lb/CODEC/Dataset/Beirut-2007-02-Viral_Hepatitis_C	Ministry of
1,046		http://linked.aub.edu.lb/CODEC/Dataset/Beirut-2007-03-Viral_Hepatitis_C	Ministry of
1,047		http://linked.aub.edu.lb/CODEC/Dataset/Beirut-2007-04-Viral_Hepatitis_C	Ministry of
1,048		http://linked.aub.edu.lb/CODEC/Dataset/Beirut-2007-05-Viral_Hepatitis_C	Ministry of

Download Merged Data (CSV)

Figure 7 Merging of Typhus and Viral Hepatitis datasets

### 3.7 Contextual Data Enrichment

This is one of the most important parts of the tool, where we made use of external knowledge graphs in order to enrich our data. Since our datasets contain Dbpedia URIs, we have developed an option through a SPARQL query that retrieves the properties of the Dbpedia URIs and adds them to the datasets. From the user's perspective, they will get add more data button, if they press on it, a SPARQL query will be sent to Dbpedia to retrieve the properties related to the columns that contain Dbpedia URIs. In other words, the existing data presented to the users in the tables based on their selection is used as a context to get additional data from Dbpedia. They will get a list of properties for them to choose from the ones they want to add for the dataset. This will facilitate data exploration and help data analysts enrich their datasets; thus, they will get better visuals.

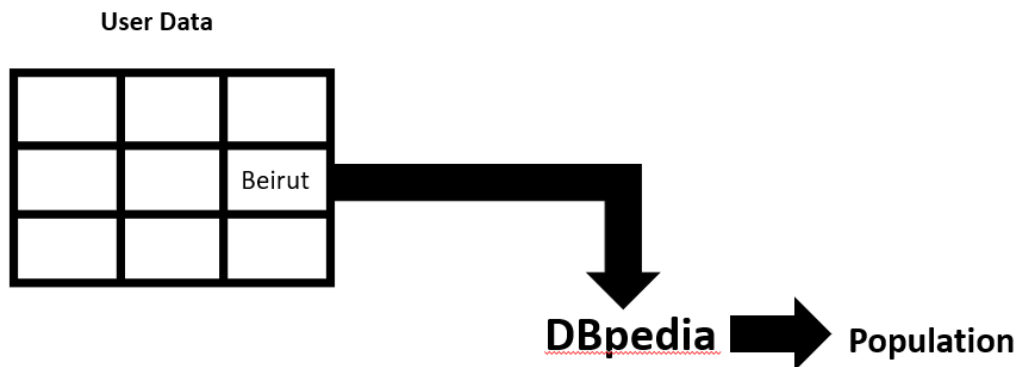


Figure 8 Illustration on Contextual Data Enrichment



# Connecting Data in the Lebanese Ecosystem

## Part of CODEC Project #OSBIMPACTS

Add more data about Area to Health\_infrastructure\_Health Status

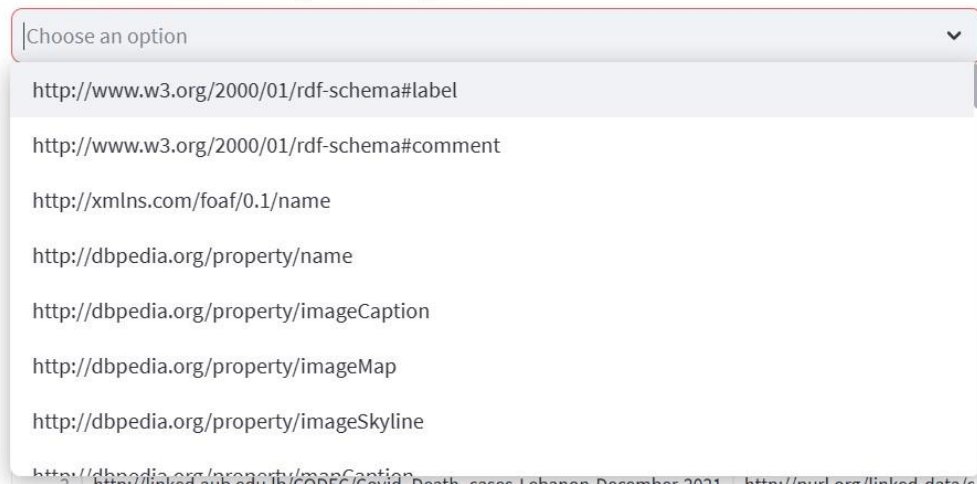


Figure 9 Data Enrichment Functionality

### 3.8 Data Download

After being done with this, users can download the csv file of each dataset separately or they can merge them in one csv file. Also, these csv files can be directly visualized on tableau to get insights.

## CHAPTER 4

### VALIDATION AND RESULTS

We validated the approach through two use cases where the tool was announced to graduate students who are working on data visualization project. The evaluation mainly consists of two parts: the quality of use cases and the interviews feedback.

#### **4.1. Boosting Tourism in Lebanon<sup>7</sup>**

Two projects have been done using our tool. The first project entitled: “Unveiling tourism patterns: A data-driven exploration of infrastructure impact”. It is done by a team of 6 data analysts who used the tool to analyze Lebanese data. They were interested in working on data related to their country and villages rather than getting data from the known open sources such as Kaggle. They discovered that our tool serves as a good database to get data about different domains related to Lebanon. Given the economic situation in the country, they decided to work in their project on data related to tourism so that they visualize it through a visualization tool. So that they can investigate the insights from the data and help in proposing a convenient solution for the problem. They aggregated 7 datasets together related to water, sewage, electricity, tourism, trade, communications, and public transportation. They then enriched their datasets by longitude and latitude factors to be able to visualize it on maps.

Users Feedback Interview data:

What they liked the most about the tool is its ability to produce for them csv files that can be visualized directly on tableau. This eased their exploration process as they did

---

<sup>7</sup> <https://sites.aub.edu.lb/datavisualization/2023/11/27/unveiling-tourism-patterns-a-data-driven-exploration-of-infrastructure-impact/>

not need to spend time cleaning and preprocessing the data. This is mainly because the data available on the tool does neither contain null values nor redundant data. They found the tool easy to deal with, also it facilitated lots of burdens in data analysis process especially in terms of cleaning, enrichment, and aggregation. Moreover, the tool contains data from different domains and categories which have eased the data collection process for us and facilitated aggregating this data. Also, we were trying to add Longitude and Latitude manually to make visuals on the maps which was time consuming for us. However, the development of the contextual enrichment functionality in the tool decreased the time needed to add data significantly, thus it enhanced the reusability of the data. Furthermore, they were happy with their ability to integrate seven datasets at the same time. They were in need for all these datasets because they all serve as indicators for the tourism factor they are analyzing such as electricity, water, communications, public transportation, and hotels.

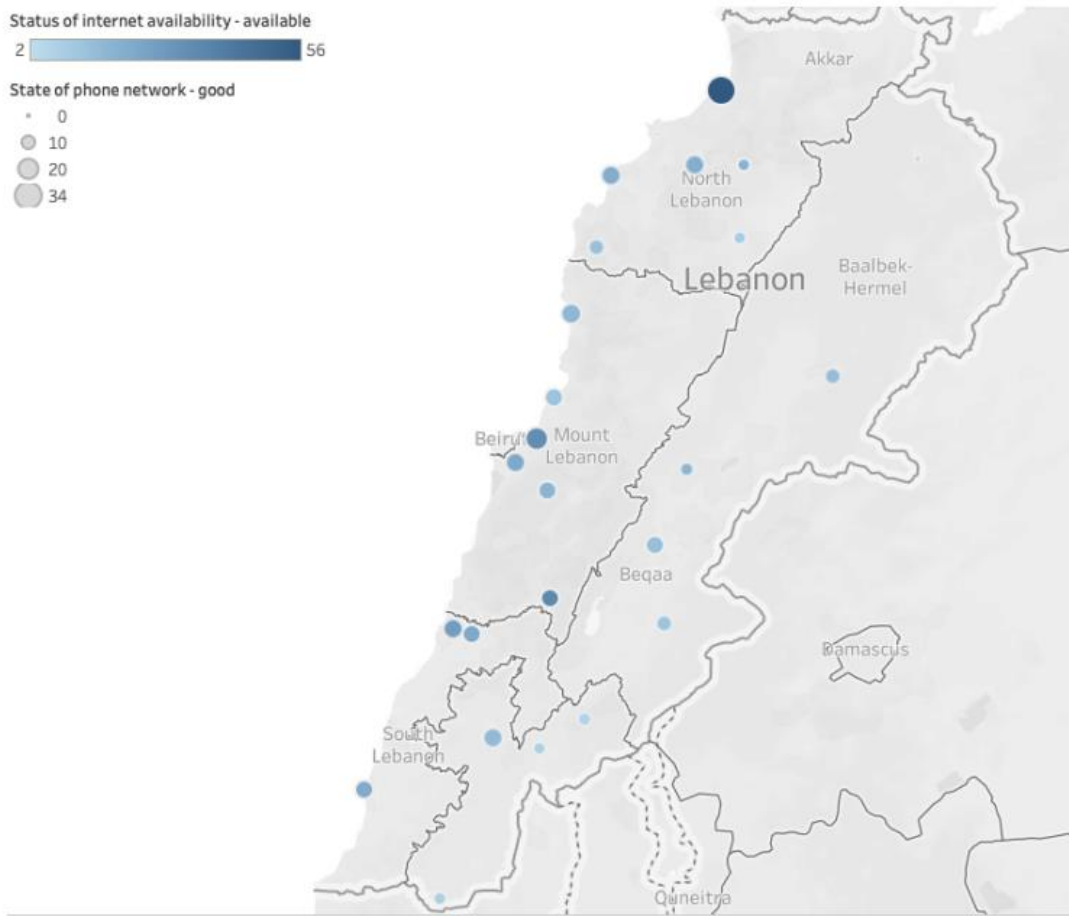


Figure 10 Status of internet availability and phone network in the Lebanese Districts and Towns

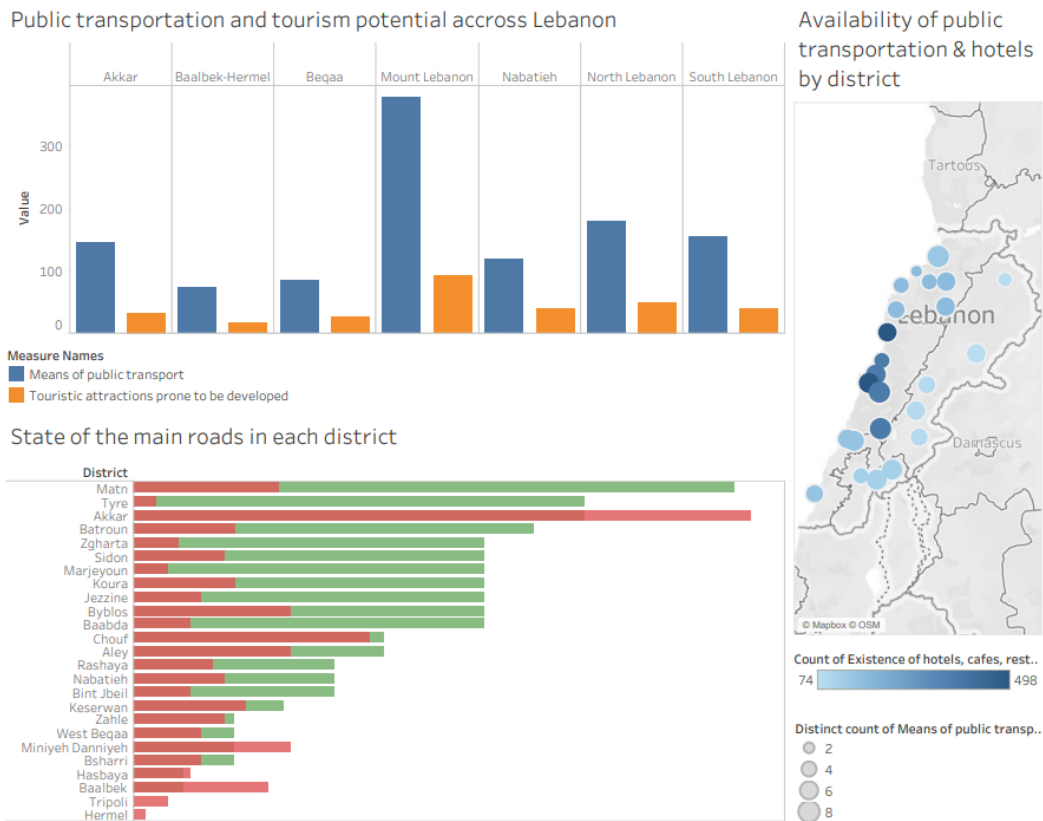


Figure 11 Dashboard of project representing Public Transportation and Hotels in Lebanon by district.

#### 4.2 Education in Rural Areas<sup>8</sup>

The second project is done by a business analyst who used our tool to investigate the educational status in rural areas and as a use case he focused on the educational data related to his home district “Zahle”. He expressed interest in analyzing data related to his village and his region to solve its problems. Also, since he was part of the problem as he was forced to move to Beirut to get his university degree because of the lack of universities in his hometown, he enjoyed dealing with this problem. Moreover, he

<sup>8</sup> <https://sites.aub.edu.lb/datavisualization/2023/11/13/education-in-zahle-district/>

enriched the data with population to investigate the number of people who are affected by the problem and the percentage of them out of the Lebanese population.

User Feedback: Interview data

He found that the generated csv files are “Tableau Friendly” since they do not need preprocessing to be visualized on tableau. Also, he claimed that the tool is easy to use and reduces the data exploration time significantly. Moreover, it made data exploration possible for beginners who have limited skills in the field.

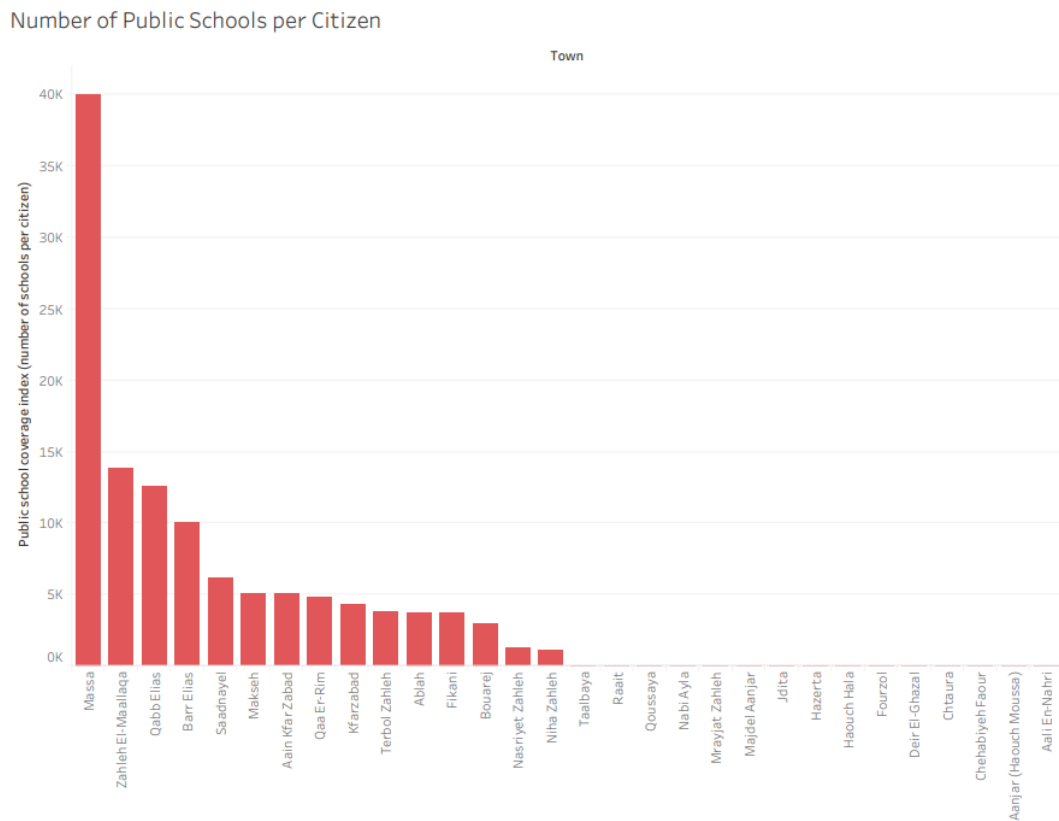


Figure 12 Number of Public schools per citizen

Percentage of Elderly (above 65)

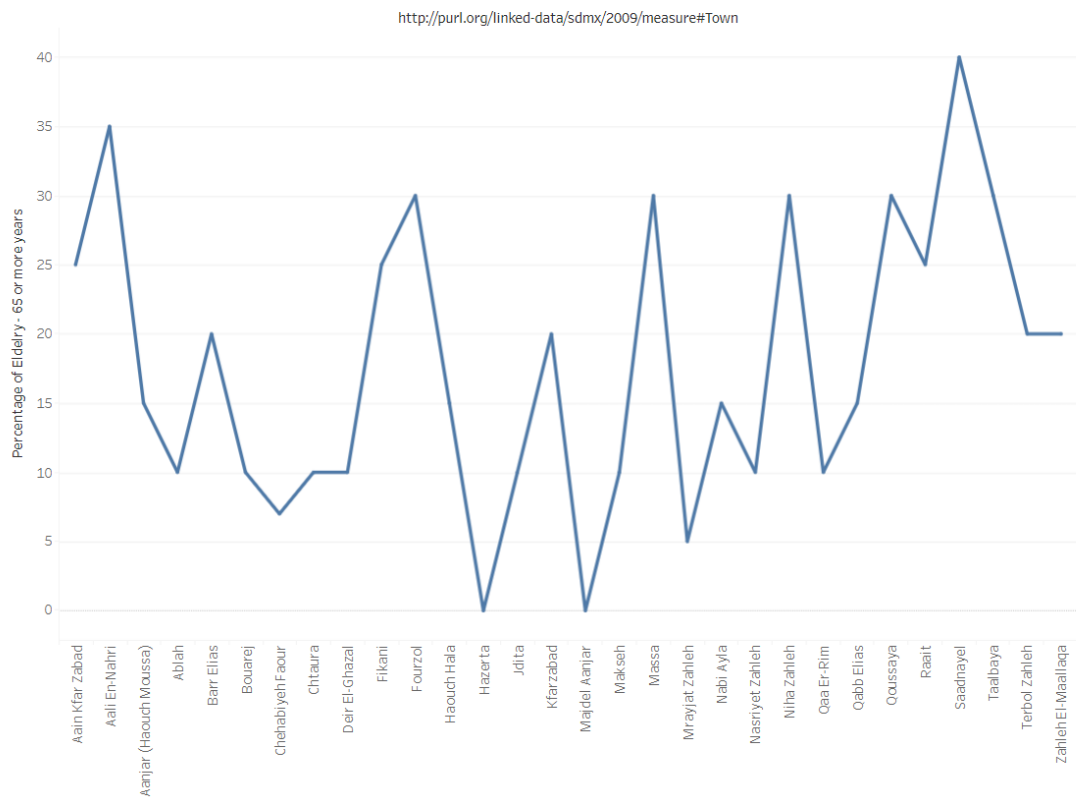


Figure 13 Percentage of Elderly in Zahle villages

## CHAPTER 5

### CONCLUSION

Overall, Data exploration is one of the key steps in the data analysis process. It involves several challenges such as data disconnection, multiple formats, inconsistency, and insufficiency. These challenges are addressed in the literature where knowledge graphs are used to connect and explore data. However, the problem of needing more data on the fly is not addressed clearly. Moreover, data analysts, especially beginners, are struggling in data preprocessing before visualizing them on visualization platforms because it is time consuming and needs programming skills. These challenges are due to having multiple data sources (as a use case here Impact, Ministry of Public Health, etc....), which leads to different formats that are hard to merge and enrich. Thus, streamlining it is vital for the sake of making the job of data analysts easier. To go over these bumps, we utilized knowledge graphs because they offer explicit semantic connections between datasets, and with external ones. These connections will offer the users the ability to contextually enrich their data on the fly. That is because the external data involves open data such as Dbpedia which is rich in data which will empower our data. Furthermore, to democratize access and usability, this solution is packaged within a user-friendly interface utilizing Streamlit. Streamlit's intuitive design and functionality aim to empower all users, irrespective of their technical expertise, to effortlessly leverage the tools and functionalities offered. Moreover, the produced datasets were ready to be visualized on exploratory visualization tools such as Tableau. Overall, KGFusionX was able to connect 78 datasets with each other by converting them to knowledge graphs where 749,500 RDF triples were produced. All these datasets can be easily enriched using their context by KGFusionX. This tool is validated through two



use cases where two data visualization projects were produced. The tool helped in maintaining high quality visuals due to the merging and enriching functionalities that helped in the reusability and exploration of the data. In these projects, data was contextually enriched from Dbpedia (such as longitude and latitude) helped in producing visuals on maps which has leveraged the value of the existing data. This research mainly contributes to streamlining the intricate processes of contextual data enrichment. This approach endeavors to transform the traditionally arduous task of data exploration into a more seamless and accessible endeavor. Also, this approach produces datasets that can be directly visualized on Tableau without need for data processing. Ultimately, this not only eases the workload of data analysts but also opens doors for a broader spectrum of users to extract valuable insights from complex datasets efficiently. Thus, the main contributions of this thesis are improving the integration of different datasets together in a clean format, contextually enriching tabular data on the fly using linked open data and creating datasets that could be easily visualized on data visualization tools such as Tableau and PowerBI.

### **5.1 Meeting the aims and objectives**

This thesis focused on meeting the following aims and objectives and they were achieved by our framework:

1. Knowledge graphs enabled connecting different datasets from different sources and domains using common dimensions (e.g. refArea, refPeriod, Disease).
2. Distributed data is converted to knowledge graphs using a flexible and dynamic code.

3. Datasets were enriched from linked open data sources (e.g. Dbpedia) where the users can choose what they want to add.
4. The integrated and enriched datasets can be visualized directly on Tableau or any other visualization tool with no need to do intensive preprocessing of the data.
5. All the above functionalities are presented in an interactive interface where users can filter, integrate, enrich, and download data then visualize it.

## **5.2 Research Limitations**

We will discuss some of the limitations of this work. Starting from data collection, there are limited data about Lebanon because most of it is not digitalized and not published online. Moreover, there are limited databases that contain data about Lebanon. This problem could be addressed by collecting data through other ways such as surveys or through academic institutions. Also, the multilingual issue limits somehow our work because of the encoding issue. Moreover, some keywords such as the towns are not all found on knowledge graphs open sources such as Dbpedia which puts another burden on our work. This could be addressed by creating additional linked data on Lebanon to fill this gap. Also, although Dbpedia is rich in properties, sometimes the users may not find the exact property they are looking for. This stimulates us to look for other linked data open sources.

## **5.3 Future Work**

Although this proposed framework has yielded promising results, avenues for future work remain open. This includes further refinement of the knowledge graph

generation, automating the process, and exploration of additional backend data sources. The proposed initiatives encompass diverse aspects of data enrichment, conversion, merging, and collaboration. One focal point involves layered data enrichment, aiming to expand and deepen datasets by exploring additional enrichment opportunities upon initial enhancements. An example of this approach involves enriching country-related data subsequent to enriching district columns with country names. Another significant endeavor pertains to real-time tabular to RDF conversion, streamlining and automating the process to swiftly transform tabular data into RDF format, expediting integration efforts. Additionally, efforts are directed towards diversifying merging techniques, seeking alternatives to traditional full outer joins to offer more flexible and efficient options for data merging. Furthermore, there's a focus on developing dataset sharing functionality, empowering users to save and share generated datasets, fostering collaboration, and allowing others to build upon existing work. These initiatives collectively aim to enhance data quality, accessibility, and collaboration within the data ecosystem.

## APPENDIX

- 1) “KGFusionX” Link: <https://interface-48gwcmgghwwg4ykmve35etj.streamlit.app/>
- 2) Project 1: “Unveiling tourism patterns: A data-driven exploration of infrastructure impact.” Link:  
<https://sites.aub.edu.lb/datavisualization/2023/11/27/unveiling-tourism-patterns-a-data-driven-exploration-of-infrastructure-impact/>
- 3) Project 2: "Education in Rural Areas". Link:  
<https://sites.aub.edu.lb/datavisualization/2023/11/13/education-in-zahle-district/>
- 4) Coding link: <https://github.com/MSBAshadi/KGFusionX>

## REFERENCES

1. Allen, M. (Ed.) (2017). The sage encyclopedia of communication research methods. (Vols. 1-4). SAGE Publications, Inc, <https://doi.org/10.4135/9781483381411>
2. Al-Tawil, M., Dimitrova, V., & Thakker, D. (2020). Using knowledge anchors to facilitate user exploration of data graphs. *Semantic Web*, 11(2), 205–234.
3. Arenas-Guerrero, J., Alobaid, A., Navas-Loro, M., Pérez, M. S., & Corcho, O. (2023, May). Boosting Knowledge Graph Generation from Tabular Data with RML Views. In *European Semantic Web Conference* (pp. 484-501). Cham: Springer Nature Switzerland.
4. Bin, S., Stadler, C., Radtke, N., Junghanns, K., Gründer-Fahrer, S., & Martin, M. (2023). Base Platform for Knowledge Graphs with Free Software.
5. Blomqvist, E., d’Amato, C., de Melo, G., Gayo, J. E. L., Kirrane, S., Navigli, R., ... & Zimmermann, A. (2021). Knowledge Graphs.
6. Bodker, S. (2021). *Through the interface: A human activity approach to user interface design*. CRC Press.
7. Bouziane, A., Bouchiha, D., & Doumi, N. (2020, December). Annotating Arabic Texts with Linked Data. In *2020 4th International Symposium on Informatics and its Applications (ISIA)* (pp. 1-5). IEEE.
8. Chen, X., Jia, S., & Xiang, Y. (2020). A review: Knowledge reasoning over knowledge graph. *Expert Systems with Applications*, 141, 112948.
9. Eberendu, A. C. (2016). Unstructured Data: an overview of the data of Big Data. *International Journal of Computer Trends and Technology*, 38(1), 46-50.
10. Ermilov, T., Moussallem, D., Usbeck, R., & Ngomo, A. C. N. (2017, August). GENESIS: a generic RDF data access interface. In *Proceedings of the International Conference on Web Intelligence* (pp. 125-131).
11. Escobar, P., Candela, G., Trujillo, J., Marco-Such, M., & Peral, J. (2020). Adding value to Linked Open Data using a multidimensional model approach based on the RDF Data Cube vocabulary. *Computer Standards & Interfaces*, 68, 103378.
12. Fafalios, P., & Tzitzikas, Y. (2019, April). How many and what types of SPARQL queries can be answered through zero-knowledge link traversal? In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing* (pp. 2267-2274).

13. García, R., López-Gil, J. M., & Gil, R. (2022). Rhizomer: Interactive semantic knowledge graphs exploration. *SoftwareX*, 20, 101235.
14. García, R., López-Gil, J. M., & Gil, R. (2023). User Experience Benchmarking and Evaluation to Guide the Development of a Semantic Knowledge Graph Exploration Tool.
15. Haase, P., Herzig, D. M., Kozlov, A., Nikolov, A., & Trame, J. (2019). metaphactory: A platform for knowledge graph management. *Semantic Web*, 10(6), 1109-1125.
16. Idreos, S., Papaemmanouil, O., & Chaudhuri, S. (2015, May). Overview of data exploration techniques. In *Proceedings of the 2015 ACM SIGMOD international conference on management of data* (pp. 277-281).
17. Ikkala, E., Hyvönen, E., Rantala, H., & Koho, M. (2022). Sampo-UI: A full stack JavaScript framework for developing semantic portal user interfaces. *Semantic Web*, 13(1), 69-84.
18. Jares, A., & Klimek, J. (2021, November). Simplod: Simple SPARQL Query Builder for Rapid Export of Linked Open Data in the Form of CSV Files. In *The 23rd International Conference on Information Integration and Web Intelligence* (pp. 415-418).
19. Johnson, J. M., Narock, T., Singh-Mohudpur, J., Fils, D., Clarke, K., Saksena, S., & Yeghiazarian, L. (2022). Knowledge graphs to support real-time flood impact evaluation. *AI Magazine*, 43(1), 40- 45.
20. Khorasani, M., Abdou, M., & Hernández Fernández, J. (2022). Streamlit Basics. In *Web Application Development with Streamlit: Develop and Deploy Secure and Scalable Web Applications to the Cloud Using a Pure Python Framework* (pp. 31-62). Berkeley, CA: Apress.
21. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., ... & Bizer, C. (2015). Dbpedia—a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic web*, 6(2), 167-195.
22. Liu, J., Chabot, Y., Troncy, R., Huynh, V. P., Labbé, T., & Monnin, P. (2022). From tabular data to knowledge graphs: A survey of semantic table interpretation tasks and methods. *Journal of Web Semantics*, 100761

23. McCusker, J., & McGuinness, D. L. (2023, May). Whyis 2: An Open Source Framework for Knowledge Graph Development and Research. In European Semantic Web Conference (pp. 538-554). Cham: Springer Nature Switzerland.
24. Mishra, P., Biancolillo, A., Roger, J. M., Marini, F., & Rutledge, D. N. (2020). New data preprocessing trends based on ensemble of multiple preprocessing techniques. *TrAC Trends in Analytical Chemistry*, 132, 116045.
25. Morsey, M., Lehmann, J., Auer, S., & Ngomo, A. C. N. (2012). Usage-centric benchmarking of RDF triple stores. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 26, No. 1, pp. 2134-2140).
26. Nuzzolese, A. G., Presutti, V., Gangemi, A., Peroni, S., & Ciancarini, P. (2017). Aemoo: Linked data exploration based on knowledge patterns. *Semantic Web*, 8(1), 87-112.
27. Patel, J. M., & Patel, J. M. (2020). Web scraping in python using beautiful soup library. *Getting Structured Data from the Internet: Running Web Crawlers/Scrapers on a Big Data Production Scale*, 31-84.
28. Quattrini, R., Pierdicca, R., & Morbidoni, C. (2017). Knowledge-based data enrichment for HBIM: Exploring high-quality models using the semantic-web. *Journal of Cultural Heritage*, 28, 129-139.
29. Rosén, G. (2019). Analysis of Tabula: A PDF-Table extraction tool.
30. Santipantakis, G. M., Kotis, K. I., Glenis, A., Vouros, G. A., Doulkeridis, C., & Vlachou, A. (2022). RDF-Gen: generating RDF triples from big data sources. *Knowledge and Information Systems*, 64(11), 2985-3015.
31. Scarpato, N. O. E. M. I., & Alessio, G. I. A. N. F. R. A. N. C. O. (2017). SAGG: a novel linked data visualization approach. *Journal of Theoretical and Applied Information Technology*, 95(22), 6192-6203.
32. Tennison, J., Cyganiak, R., & Reynolds, D. (2012). *The rdf data cube vocabulary*. Technical report, W3C Working Draft 05 April, 2012. <http://www.w3.org/TR/vocab-data-cube>.
33. Tianyi, Z., Yun, S., & Lee, Y. (2020). Performance Comparisons of Three SPARQL Query Types for Fuseki and AWS Storage Platform. *한국정보과학회 학술발표논문집*, 136-138.

34. Thalhammer, A., Lasierra, N., & Rettinger, A. (2016). Linksum: using link analysis to summarize entity data. In *Web Engineering: 16th International Conference, ICWE 2016, Lugano, Switzerland, June 6-9, 2016. Proceedings 16* (pp. 244-261). Springer International Publishing.
35. Tyagi, S., & Jimenez-Ruiz, E. (2020). LexMa: Tabular data to knowledge graph matching using lexical techniques. *CEUR Workshop Proceedings, 2775*, pp. 59-64.
36. Tvarožek, M., & Bieliková, M. (2010). Generating exploratory search interfaces for the semantic web. In *Human-Computer Interaction: Second IFIP TC 13 Symposium, HCIS 2010, Held as Part of WCC 2010, Brisbane, Australia, September 20-23, 2010. Proceedings* (pp. 175-186). Springer Berlin Heidelberg.
37. Vogt, L. (2023). Extending FAIR to FAIREr: Cognitive Interoperability and the Human Explorability of Data and Metadata. *arXiv preprint arXiv:2301.04202*.
38. Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3(1), 1-9
39. Yulianti, G., Krisnadhi, A. A., & Hilman, M. H. (2021, October). Transforming Table to Knowledge Graph using A Rule-based Pipeline. In *2021 International Conference on Advanced Computer Science and Information Systems (ICACSIS)* (pp. 1-10). IEEE.