TRANSFORMING OPEN DATA TO LINKED OPEN DATA: AN ONTOLOGY FRAMEWORK BASED ON NEW ZEALAND CASE STUDY

A THESIS SUBMITTED TO AUCKLAND UNIVERSITY OF TECHNOLOGY
IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Supervisor

Dr. Parma Nand

Associate Prof. Minh Nguyen

June 2022

By

Paramjeet Kaur

School of Engineering, Computer and Mathematical Sciences

Abstract

Open government initiatives are increasingly gaining momentum and are becoming a critical part of the democratic fabric of both developing and developed nations. The primary motivation behind the initiatives is to provide transparency to the general public and encourage wider data usage by not only government-employed professionals but also by the general public and data scientists. As part of open governance, governments around the globe are increasingly releasing governance-related data that was either classified or undisclosed before the open data initiative. However, the consumption of the data has been dearth largely due to the nature of the data being disparate and heterogeneous formats. This has given rise to a need for frameworks that would be able to transform this data into an easily consumable form and make it accessible to stakeholders as well as the general public. This thesis proposes the design, implementation, and usage of an approach driven by an ontology that captures the knowledge of a subset of data released by the New Zealand government as a part of the open data initiative. The proposed framework transforms open data into linked open data based on a novel ontology developed as part of this research.

While open data is commonly available in huge quantities, it lacks quality, accuracy, consistency, and completeness. It can be challenging to find information from this data for analysis towards an objective. There are many rich open data repositories globally. However, they are challenging to understand and use, because the data can only be accessed with a complex set of key phrase search options. Even then, it might

end up retrieving data that might be irrelevant and/or incomplete. To mitigate this, ontology-based search, which uses semantics rather than keywords, has been proven to be more effective in strengthening the quality of queries for searching for content from repositories.

This thesis presents a novel framework for semantically linking and achieving disparate open datasets. The framework and end-to-end process are demonstrated using open datasets for agriculture, land, and rainfall sectors in New Zealand. The framework is used to generate ontologies, which are then populated using the data and stored in a knowledge base. We then demonstrate how the knowledge base can be used to extract valuable, rich information pertaining to an objective. We demonstrate how ontologies can be linked manually as well as semi-automatically. Manual linking requires domain experts, whereas semi-automatic linking reduces the overhead of the dependency on domain experts to manually link the concepts. The result of this approach is promising in terms of enhancing the quality of data and the efficiency of the search.

An expert evaluation was conducted to demonstrate and evaluate the efficiency and effectiveness of the ontology framework. The proposed framework was given to seven domain experts to access the knowledge base and do an end-to-end evaluation. The evaluators were asked to answer questions on five criteria: usability, reliability, correctness, usefulness, and effectiveness. A thematic review was then conducted for the collated feedback of domain experts using Nvivo. The results demonstrate the proposed scheme can effectively link open data by generating ontologies for disparate open data, which can then be used to supply useful information derived from a conflation of the original data sets.

Contents

List of Tables

List of Figures

Attestation of Authorship

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the qualification of any other degree or diploma of a university or other institution of higher learning.

Signature of candidate

Co-authored Work

All co-authors in the following table have approved these chapters for inclusion in Paramjeet Kaur's doctoral thesis.

Chapter	Author %
Chapter-2 Paramjeet Kaur, Parma Nand Towards Transparent Governance by Unifying Open Data. Manuscript Published in IAENG International Journal of Computer Science (Scopus Indexed)	PK- 85% , PN- 15%
Chapter-3 Paramjeet Kaur, Parma Nand, Salman Naseer, Akber Abid Gardezi, Fawaz Alassery, Habib Hamam, Omar Cheikhrouhou, Muhammad Shafiq Ontology-Based Semantic Search Framework for Disparate Datasets. Manuscript published in Intelligent Automation and Soft Computing	PK- 80% , PN- 6% , SN- 3% , AAG- 2% , FA- 2% , HH- 2% , OC- 2% , MS-3%
Chapter-4 Paramjeet Kaur, Parma Nand Evaluation of New Zealand Open Data Ontology Framework based on Experts Knowledge. Manuscript Submitted in IAENG International Journal of Computer Science (Scopus Indexed)	PK- 85% , PN- 15%
Paramjeet Kaur (PK), Parma Nand (PN), Salman Naseer(SN), Akber Abid Gardezi (AAG), Fawaz Alassery(FA), , Habib Hamam(HH), Omar Cheikhrouhou (OC), Muhammad Shafiq (MS)	

We, the undersigned, hereby agree to the percentages of participation in the chapters identified above

Paramjeet Kaur	Parma Nand	Salman Naseer	Akber Abid Gardezi	Habib Hamam	Muhammad Shafiq	Fawaz Alassery	Omar Cheikhrouhou

Acknowledgements

I would like to convey my heartfelt gratitude to my primary supervisor, Dr. Parma Nand, for his patience, encouragement, and extensive knowledge in supporting my PhD studies and research.
His guidance was tremendously helpful during my PhD studies and thesis writing. He not only taught me the research strategies and methods, but also encouraged and motivated me on occassions when I was feeling low. For my PhD study, I could not have imagined having a better supervisor.
I am also grateful to my secondary supervisor Associate Prof. Minh Nguyen for his guidance and support.
Completing this research would have been difficult without the support received from Auckland University of Technology (AUT). I would like to acknowledge the School of Engineering, Computer and Mathematical Sciences (SECMS) and faculty of Design and Creative Technologies for their invaluable support.
I wish to express my gratitude to my father Dhir Singh, and mother Tejwant Kaur for their love, care and support throughout my studies. My lovely husband Jatinder Singh, who has always encouraged and motivated me to go above and beyond, deserves special mention.
Lastly, I would like to thank my friends Lankesh Weerasekara, and Amr Van Den Adel for their assistance during my research.

Dedication

To my husband Jatinder Singh whose encouragement, patience, and passion helped me to accomplish this research work.

Chapter 1

Introduction

1.1 Introduction

This chapter introduces the research topic and the factors that motivated the researcher to conduct the proposed study. In addition, the chapter addresses the gaps in the existing literature and then describes the gaps that gave rise to this research. Lastly, at the end of this chapter, the thesis structure is presented.

This thesis presents a framework that can be used to systematically transform disparate open data into searchable, linked open data. We demonstrate the end-to-end framework using New Zealand (NZ) open datasets for agriculture, land, and rainfall sectors. The framework incorporates various steps required to integrate raw data released by various government departments and other sources that might not be directly related. The framework specifies the steps required to transform the raw data from the formats in which they were released by semi-automatic definition of the semantic links between the datapoints and transforming the datapoints into RDF triples, which can then be archived in a knowledge base accessible via SPARQL or another type of endpoint. The framework presents the essential stages required using a case study, but more importantly, it provides a catalyst for further research and extension of the framework

to cover the wide ranging data being made available as part of the open data momentum. This framework brings together different data processing techniques and theoretical data science to organise and annotate knowledge in a way that makes sense.

The first section of this chapter provides context and motivation for the measures that led to the conduct of this research. The research problem is discussed in section 2, followed by the research questions in section 3. Sections 4 and 5 highlight the research contributions and publications, respectively, followed by the thesis structure in Section 6.

1.2 Background

Open Government Data (OGD) is described as non-confidential and non-private data produced or managed by a government with public funds and released publicly without limitations on usage or distribution (Janssen, Charalabidis & Zuiderwijk, 2012). Open data has seen substantial development in the last few years. Many countries are initiating the paradigm of open data and opening up government data to the public. From a technical point of view, open data refers to any freely available datasets. There is no requirement to purchase a patent or licence to use the data. The data is available and can be used and accessed by researchers, universities, and other groups of people from the web community (Harrison & Sayogo, 2014).

Datasets supplied by the government under open data standards contain a wide variety of information that can influence the decisions and activities of stakeholders spanning from people to companies (DiFranzo et al., 2011). Over the last few years, open data (OD) usage has grown in popularity among government agencies, businesses, and citizen groups. The principle of open government is that a citizen has the right to obtain public information, data, and processes freely. The availability of open data has numerous benefits for the public and other stakeholders, (Fleiner, 2018). The potential

use of open data can be illustrated with the following examples:

The centres for Medicare services in the United States publish data on the quality and facilities provided by each hospital and rest home in the country that acknowledges Medicare. If represented correctly, this data could allow individuals to make better healthcare decisions by comparing the hospital with other medical facilities. However, open government initiatives publish raw, disparate data in heterogeneous formats, making it challenging to link and use the data without an arduous effort.

When anyone can access and allocate data freely, it is known as "open data". However, open data is not equivalent to or identical to linked data. As per the definition of open data, it can be accessible to anyone even without linking the datasets. Simultaneously, data can be linked without making it freely available. Hence, linked open data is a potent combination of open and linked data as it is associated and freely available (Fleiner, 2018). A well-known example of linked open data is DBPedia. Linked data, or linked open data, is a technique of data publication that utilises popular web technologies to connect and access associated information on the web. It focuses primarily on finding resources with HTTP (Hypertext Transfer Protocol), URIs (Uniform Resource Identifiers), and using norms like RDF (Resource Description Framework) by offering details on such resources and linking them with other web-based resources (Mekhabunchakij, 2016).

Linked data has become a new research area that can be employed as a technique for representing complex data. The word "linked data" is also used to describe a collection of protocols to publish and connect structured data on the web. It deliberately promotes the use of dereferential links, utilising linked data concepts. It is deployed by multiple types of technologies, such as RDF and XML. Search engines are available that enable users to crawl through this data web and execute user query outcomes (Bizer, Heath & Berners-Lee, 2011). Linked data is a practical method for publishing structured information on the Web to find relevant information from various sources. Using URIs,

in a nutshell, allows us to refer in an absolute way to concepts and things, whether real or imaginary. To make persistent use of things, Tim Berners Lee suggested that a URI should be provided for any resource of relative importance. By using the commonly adopted http: URI system, the linked data concept begins to provide a representation of required resources using http not only for linking files, but also for linking information across the internet. The design scheme of Linked data is focused on an open world assumption and utilises dereferable HTTP URIs to identify and access information objects, RDFs to define the metadata of those items, and semantic connections to define the interactions amid those items (Refaeilzadeh, Tang, Liu, Liu & Özsu, 2009).

The heterogeneous format of open data makes it challenging to parse and understand. To make it meaningful, it should be processed in line with an objective (Attard, Orlandi & Auer, 2016). For instance, if a country publishes the unemployment rate of each region by interlinking the unemployment data with health or hospital data, the quality of health of the individuals can be investigated in that particular region. This interlinked data can help identify unemployed individuals' health status. Moreover, the quality and services of the hospitals can also be investigated. As demonstrated in the preceding example, a single dataset can be used to support individuals, companies, and communities in multiple ways. By linking datasets, a user can conceive the information by looking at other data sources. For example, if a new entrepreneur wants to buy a business and is unsure how to select an appropriate business, that will bring more profit to him. Linked open data (LOD) could aid in the decision-making process of choosing the right business. To achieve this, the data from business and financial sectors would be linked using LOD principles, which will graph the progress of the business and articulate the year in which the company has reached a significant success rate. Thus, end-users would be able to make more informed decisions.

Moreover, LOD will provide transparency to the general public so that individuals can get to know what is happening in the country. For instance, by making the annual

financial spending data available as LOD, the public will have access to the data and quickly figure out where the money is being invested. As a result, government bodies need to justify their operations by providing data sets that will give a clear picture of all spending and decrease the chances of bankruptcy and embezzlement. Hence, there is a need to have an approach to assist in the interlinking of diverse datasets where semantic encoding can be achieved.

A growing number of countries understand the need for transparency in governance. Several governments have been attempting to make government operational data more accessible to the public.Yet, the data is vast, heterogeneous, and segregated, making it difficult for the general public and information consumers to use. Several countries have engaged in open data programmes, attempting to link data using a variety of frameworks and publishing methods (Kaur & Nand, 2021b).

Recent efforts to open government data are rapidly gaining popularity. Although it provides enormous advantages for improved clarity, the issue is that the information is often available in diverse formats, lacking simple semantics that explain the data reference (Nikiforova & McBride, 2021) . Moreover, data is also presented in ways that a broad spectrum of user groups who need to take essential decisions cannot understand clearly. Due to the heterogeneous nature of the data, it is challenging to integrate it as it reduces the possibilities of information sharing (Klein, Klein & Luciano, 2018).

The majority of the research thus far has focused on the usage and availability of open data via data portals and on examining and comparing the quality of the portals (Charalabidis, Alexopoulos & Loukis, 2016), (Máchová & Lněnička, 2017), (Altayar, 2018), (Quarati & De Martino, 2019). There is a dearth of studies on specific frameworks, guidelines, and models to support open data access in a semantically integrated context. Therefore, there is a need for such a framework that can semantically link the diverse data sources and can make data accessible and more easily useable.

1.3 Research Problem and Motivation

Open data supports economic growth and the establishment of new enterprises (Foundation, 2018). It has a lot of potential and utility, so it's regarded as a critical resource and primary material for a wide range of innovative products and services (Petrov, Gurin & Manley, 2016). In recent times, many open data sets are available online, either in structured or unstructured formats. For instance, the NZ government has established a data portal, which gives access to all open data sets of various sectors, including but not limited to agriculture, land, education, energy, environment, marine, forest, and rainfall. The global data index of 2014 shows that 72% of New Zealand data sets are available as open (Index, 2014). Additionally, countries such as the United Kingdom, Australia, Denmark, France, Finland, Norway, the United States, Germany, and India rank among the top 10 countries with the maximum open data percentages. The availability of open data has several valuable benefits for the general public and professionals who use data for making decisions and planning purposes. Encoding OGD as linked open data (LOD) would enable a user to browse a data source and then navigate the links into other related data sources to get all the relevant data in one place (data.govt.nz, 2020).

With enormous quantities of data accessible on the web, discovering the data and observations of interest becomes an important and challenging problem. One such challenge comprises the capacity to find data on the web that is useful and relevant to a user or application (Patel & Jain, 2021). The interlinking of OD into LOD allows the utilisation of data within organisations and domains such as statistics, research, science, health, education, publications and more. By connecting data sources, interrelationships and associations can be rapidly recognised. The segments of data and information can be organised, interchanged, exported, and linked with the help of the Uniform Resource Identifier (URI) and Resource Description Framework (RDF). This technique allows the interlinking of free storage data from various sources without constraints on use and

composition (Kalampokis, Tambouris & Tarabanis, 2011).

While open data has several benefits, it also entails numerous technology, legislation, use, and complexity barriers. Open data does not hold any value on its own. It only becomes useful and valuable when utilised in an application for some objective, primarily for decision support. There is an abundance of misconceptions about open data. For example, all information should be published freely, and publishing open data will automatically bring transparency. These misconceptions are used to convince data providers to open up their data to the public. However, these do not take into account the various limitations and the heterogeneous nature of open data. Due to the heterogeneity of formats and information sources, managing and utilising open government data is challenging. The gap can be identified as that there is no specific framework, guidelines, or models to support the full access and semantic integration of open data. Therefore, there is a need for a framework which can semantically link the diverse data sources and can make open data accessible and useable. The primary motivation behind this research is to build a process and framework that can transform these datasets into a knowledge base from which useful information can be extracted with minimal effort by professionals, data consumers, as well as the general public. We present an approach that can provide quality linked data ontology from New Zealand Government data sources on agriculture, land, and rainfall sectors. The main feature of our framework is that it allows for stable RDF (Resource Description Framework) data that can be improved without interrupting existing applications. It will be seamlessly consumable by the stakeholders.

1.4 Research Significance

The significance of this research lies in the fact that it will enable the interlinking of diverse datasets here with useful semantic encoding. The result of the study will be of

great benefit in the following areas:

Transparency: It will promote transparency for the government. It will support accountability as citizens can keep an eye on the basis for decision making for the government, and they can question the government on decisions. It will help citizens find and use the information freely and openly.

Reusability: Data openness makes it available for the developers to reuse and build valuable applications or websites that can contribute to the country's growth.

Economic Growth: Entrepreneurs can use linked open data to build novel or innovative business ideas and products that will contribute to the country's economic growth and stimulate development.

Public Participation: Linked open data enhances the public's engagement, which indirectly helps the government's effectiveness and decision-making. The government will inform the citizen about their actions. This will help to build trust between the government and citizens. It will also improve the process and services of the government.

1.5 Research Scope

The latest efforts to open government data is rapidly gaining popularity. Although it provides enormous advantages for improved clarity, the issue is that the information is often available in diverse formats, lacking simple semantics that is able to explain the nexuses and nuances between entities in the data. Moreover, data is also presented in ways that is not conducive to a broad spectrum of user groups who need to make crucial decisions based on an understanding of the data. Due to the heterogeneous nature of the data, it is challenging to integrate it as it involves linking entities with semantic relations which might outwardly seem to be only remotely linked. This thesis presents an ontology framework to transform disparate open data into linked open data. The

scope of this thesis encompasses a subset of data released by the NZ government for the rainfall, agriculture, and land sectors. The framework, on the other hand, applies to all of the data that has been released by the government of New Zealand as well as other governments around the world.

1.6 Research Questions

To design the framework's schemes, the following research questions are addressed.
RQ 1: Can disparate structured data published by various government departments be computationally with RDF encoding and semantically linked in an ontology framework?
RQ 2: Can semantically linked data with RDF encoding be made available using SPARQL endpoints to satisfy the requirements of a wide range of stakeholders and the general public?

This research aims to design an ontology framework to transform open data into linked open data and make it available via a SPARQL interface so that stakeholders and the general public can more easily consume it. Research Question (RQ) 1 focuses on the design approach and factors for transforming open data into a knowledge base. RQ 2 aims to publish the heterogeneous data generated in RQ 1 via the SPARQL interface to extract valuable information.

1.7 Research Contribution

Firstly, this thesis presents a detailed framework to transform disparate, heterogeneous data into end-user Linked Open Data (LOD) SPARQL endpoints. As a case study, we used the framework to transform the open government data extracted from the New Zealand government portal into a knowledge base accessible via the SPARQL endpoint. The framework is organised using four layers: data conversion, RDF-based ontology

generation, semantic link generation, and SPARQL interface. The overall architecture is designed in layers to facilitate flexibility, maintainability, and scalability.

The layered architecture also fosters efficiency, reliability, and usability. Unlike traditional approaches, a layered architecture enables the entire system to grow and expand as independent modules. The whole design extends and stays synchronised because the frameworks can adjust freely based on the needs of each module.

The primary reason behind the layered design was to enable the reuse of the various components of the framework by multiple applications. For example, some external systems can use only the ontology generation module, which is easily accomplished due to the architecture's ontology layer's stand-alone nature. The main contributions of this thesis are outlined below:

Contribution 1: The proposed framework has resolved the heterogeneous data problem in the data conversion and RDF-based ontology generation layers. The data conversion process is automated and takes data as input from diverse data sources.

In the first phase of the framework, either the download file or URL of the open datasets can be used as input. The data conversion process then parses the raw open datasets into comma-separated value (CSV) format. Several libraries, dialect descriptions, and processes are utilised to parse the raw open data sets.

In the second phase, the resultant parsed file is used to generate the web ontology language triples. These triples are further employed to generate ontology. Several processes, metadata annotations, and libraries are used to generate the ontology. The resulted ontology can be visualised by using the available visualisation tools such as protégé. At this point, we have shown that the proposed method can generate an ontology for any open data set.

Contribution 2: The key contribution of this work is the solution it provides for semantic link generation. We designed a semi-automatic process to generate the semantic links between two or more ontologies based on common properties. These

properties are identified manually. One particularly notable feature of this process is that where no common properties are identified, the data is still added to the semantically linked ontology knowledgebase to prevent data loss.

Contribution 3: We developed a SPARQL interface to publish the disparate data sources consumed by stakeholders and the general public through imposed queries. The interface is designed to be user-friendly so you can select the desired ontology files easily. Depending on the choice, a user can either question a single ontology or the semantically linked ontologies.

Contribution 4: We developed an ontology framework to transform open data into linked open data based on a New Zealand case study. The proposed ontology framework is the key contribution of the research. The data sets from the agriculture, land, and rainfall industries of New Zealand are used to conduct an in-depth examination of the ontology generation and RDF semantic encoding process. To get useful information out of the ontology, a SPARQL interface is used as the endpoint. In addition, the proposed framework's efficacy and efficiency are assessed through expert evaluation.

1.8 Publications

Journal Publication

- Kaur, P., & Nand, P. (2021). "Towards Transparent Governance by Unifying Open Data". IAENG International Journal of Computer Science, 48(4).

- Kaur, P., Nand, P., Naseer, S., Gardezi, A. A., Alassery, F., Hamam, H., ... & Shafiq, M. (2022). "Ontology-Based Semantic Search Framework for Disparate Datasets". Intelligent Automation and Soft Computing, 32(3), 1717-1728.

Conference Publication

- Kaur, P., & Nand, P. (2021, April). "Implementing Automatic Ontology Generation for the New Zealand Open Government Data: An Evaluative Approach". In International Conference on Advances in Computing and Data Sciences (pp. 26-36). Springer, Cham.

Under Review Articles

- Kaur P, & Nand, P. "Evaluation of New Zealand Open Data Ontology Framework based on Experts Knowledge". International Journal of Semantic Computing (IJSC) (Submitted)

1.9 Theoretical Framework

This thesis uses a scientific methodology by defining a problem and developing a prototype as a case study on a limited scope, evaluating it, and then asserting that the framework applies to a broader scope. We conducted a study to analyse New Zealand's open data set and convert it into linked open data to extract useful information for novice users and society. We develop a prototype that generates ontologies of the open government datasets. It uses a semi-automatic approach where open datasets of agriculture, land, and rainfall sectors are used to generate semantically linked ontologies. We design a SPARQL interface to impose queries on the generated ontologies so that the data can be extracted more easily.

1.10 Thesis Structure

The proposed ontology framework transforms the open data into linked open data. We used the three sectors' open data sets as a case study to generate the ontologies and imposed SPARQL queries to extract valuable data based on the needs of the user. The

first research question "Can disparate structured data published by various government departments be computationally with RDF encoding and semantically linked in an ontology framework?" is covered in chapter 2, where a systematic ontology framework is generated using disparate data sets published by the New Zealand government.

Moreover, the SPARQL endpoints part of the second research question "Can semantically linked data with RDF encoding be made available using SPARQL endpoints to satisfy the requirements of a wide range of stakeholders and the general public?" to validate the framework against the imposed SPARQL queries is covered in chapter 2. Chapter 3 covers the first research question which describes the semantic link generation using the RDF encoding and the development of the SPARQL endpoint part of the second research question. Chapter 4 validates the whole framework, where expert evaluation is conducted to analyse the correctness, efficiency, effectiveness, and usefulness of the proposed framework.

This thesis contains five chapters, as depicted in figure 1.1. The first chapter gives an outline of the proposed study. Chapter 2 presents a review of literature, which begins with an introduction to Open Government Initiatives, the role of linked open data, and various mechanisms to connect and collate the data. The other sections of this chapter provide an in-depth literature review of multiple techniques, tools, algorithms, models, and frameworks used to transform open data into linked open data.

Furthermore, chapter 2 discusses the benefits and challenges of linked open data. Open research challenges motivate the identification of research gaps for transforming open data into linked open data. Finally, the proposed design architecture is discussed, and a case study is conducted to validate its usage, followed by a discussion and conclusion.

Chapter 3 presents the methodology used to generate the proposed framework to transform open data into linked open data. This chapter has focused on the semantic link generation and SPARQL interface, where sample SPARQL queries are used to test

the validity of the created ontology knowledgebase.

Chapter 4 presents the evaluation results of the proposed framework. An expert evaluation is conducted where industry experts use and analyse the proposed system. The experts were given a set of questionnaires to answer. Qualitative feedback from experts is analysed and processed by NVivo so that relevant themes can be found. These themes can be used to determine the framework's effectiveness, correctness, usability, and appropriateness, as well as its usability and effectiveness.

Chapter 5 concludes the research by summarising the benefits and limitations of the proposed framework with recommendations for future research.

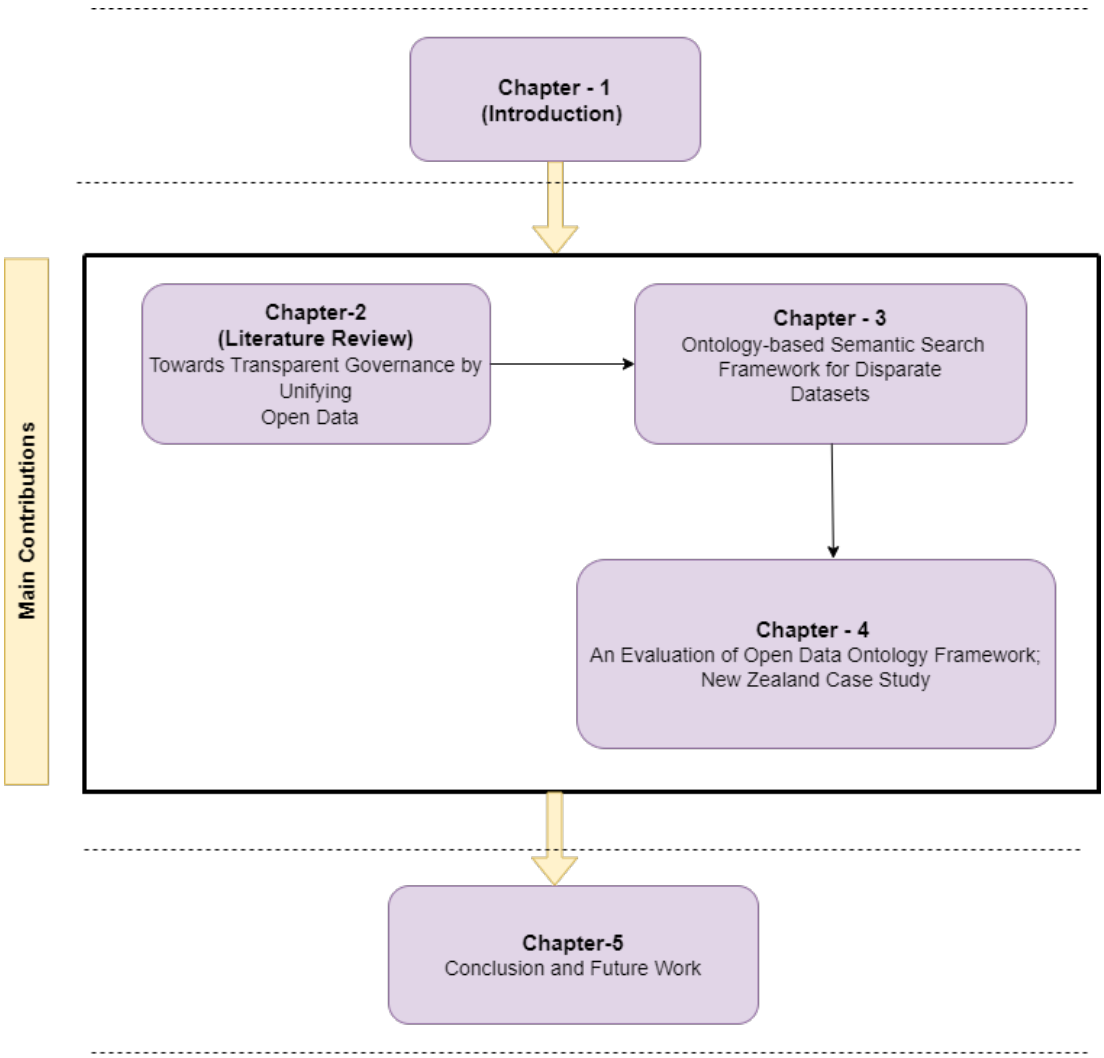


Figure 1.1: Thesis Structure

1.11 Literature Review

Chapter 2 covers the overall literature review of the framework as well as different concepts and terminologies related to open data are discussed along in various proposed studies, prototypes, architecture, and ontology frameworks of open data. Based on the literature presented and analysed in chapter 2, it is evident that existing prototypes, architecture, algorithms, tools, and ontology frameworks are designed and developed

with specific features to support certain functionalities and areas. Due to the particular functionalities, these proposed solutions cannot deal with all types of open data available worldwide.

Chapter 3 covers the literature review in the area of semantic link development. The latest developments in this area have used open government data from different countries and departments to investigate prototypes, E-GIF ontologies, search engines, and ontology-based frameworks. Although extensive research has been conducted, there is still more room to implement mechanisms and technologies to take advantage of open data procedures to extract valuable information for the benefit of the public. There is an urgent need for a new method, especially to convert different open data sources into a standard form so that a vast knowledge base can be created, and multiple data sources can be used to generate semantically rich data. The purpose of this research is to use the open government data set of the New Zealand government to develop a simple and accurate method to generate semantically rich automatic ontologies.

Chapter 4 covers the Literature review related to evaluation methods available and proposed by the scientific community in the field of ontology evaluation. Ontology evaluation is problematic due to the descriptive structure of ontologies and their use and expansion beyond a centralised monitoring mechanism. Even though there are numerous methodologies and tools available, there is no standard mechanism for ontology evaluation because ontologies are semantic and hence require human evaluation. After analysing all the evaluation techniques, it became clear that there is no single best and complete method or approach for ontology evaluation. The choice of method depends on the evaluation purpose, what aspects of ontology we are trying to test and the application in which ontology is to be used. Except for a few exceptions, most tools have been designed as plugins for desktop applications. They compute metric values but do not relate them to the criterion under consideration. After thoroughly examining all available evaluation methods, human or expert evaluation was deemed

an appropriate solution for analysing our proposed ontology framework. Figure 1.2 highlights the conceptual framework of the thesis.

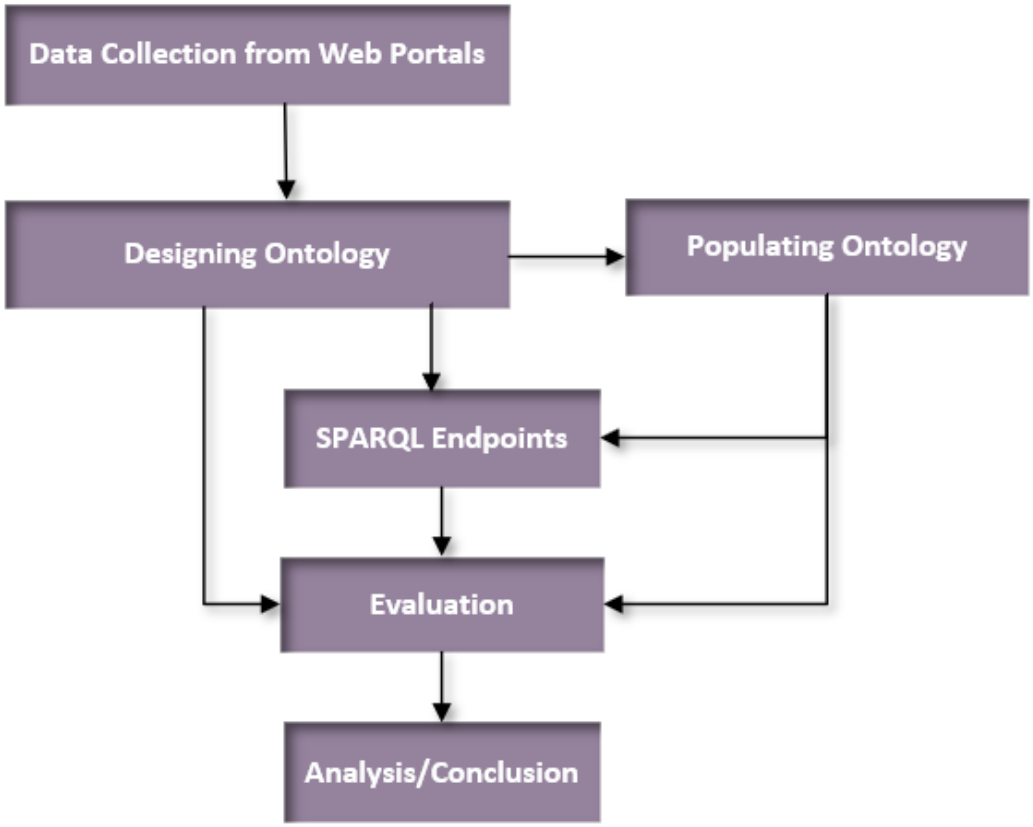


Figure 1.2: Conceptual Framework

1.12 Introduction to Chapter-2

Chapter 2: "Towards Transparent Governance by Unifying Open Data" is a published journal article and used as a chapter in this thesis. The main aim of this chapter is to present the concept of open data initiatives, and the encoding of open data as linked open data. It presents a detailed literature review of the datasets, evaluation methods, and techniques used to transform open data into linked open data. The benefits and

challenges of open data are discussed in detail. Furthermore, the architecture of the proposed framework is presented and followed by a case study evaluation using NZ open data sets of three sectors. The results are analysed and discussed.

1.13 Introduction to Chapter-3

Chapter3, "Ontology-based Semantic Search Framework for Disparate Datasets" is a published journal article and used as a chapter in this thesis. This chapter aims to discuss the overall architecture, semantic link generation process, and SPARQL interface of the framework in detail. The details of the semantic link generation process are presented. Several sample SPARQL queries are imposed to generate the results from the semantically linked ontologies pertaining to an objective. The results of the queries are analysed and discussed in the chapter.

1.14 Introduction to Chapter-4

Chapter 4, "An Evaluation of Open Data Ontology Framework; New Zealand Case Study" is a submitted journal article and used as a chapter in this thesis. The main aim of this chapter is to evaluate the proposed framework. An expert evaluation is used to analyse the usefulness, correctness, effectiveness, and efficiency of the proposed framework. The proposed framework is given to seven domain experts to evaluate. A series of questions are given to the expert to answer, which highlights the different capabilities of the framework. All of the participants demonstrated a clear understanding of the questionnaire. The qualitative feedback provided by experts is analysed and processed by using Nvivo. Appropriate changes were made to the framework based on the expert feedback.

In summary, the work presented in Chapters 2, 3, and 4 is the main research

contribution of this thesis. These chapters are discussed and presented in detail in the subsequent sections of this thesis.

Chapter 2

Towards Transparent Governance by Unifying Open Data

Abstract

Open data initiatives have been gaining increasing momentum in recent times, both with national governments as well as regional governing bodies.An increasing number of governments are realising the need for the role of transparency in governance. In congruence to this, a number of governments have been working towards opening up government operational data to the public, however the data is extremely large, disparate, and segmented, hence it is hardly useable by the general public as well as data consumers. A number of countries have worked on open data initiatives and attempted to link the data using diverse frameworks and publishing tools. This research presents an analysis of the works that have developed ontologies and frameworks to link the data released by governments as part of open data initiatives. It also provides a critical evaluation of the techniques used and suggestions for novel techniques that can be used to improve the frameworks and ontologies. We also present the results of a case study that used an ontology-based linking of data released by governments as part of the open

data initiative. The results show that an ontology mapping of such raw data drastically enhances the usability and quality of the raw data.

2.1 Introduction

The world Wide Web has been able to connect the world through the use of hyper links between the web documents. These hyperlinks are used to navigate between html pages containing information as free text. In this way, all of the documents can be accessed via a single web link, which integrates other links embedded in pages. The accelerated growth of an informed society means people are growing in awareness of their rights and they want to know, and get increasingly involved, with governance mechanisms. This has given rise to the necessity of having transparency by governing bodies, resulting in the release of data to the public based on the rationale of making policy decisions transparent. The open data model proposed by (Fleiner, 2018) highlights the benefits and use of open data in research, which can help researchers make data driven decisions. The fundamental principle of open government is that the general public has the privilege of accessing data, information, and activities generated by government agencies. The raw data, without any interpretation, plays a vital function in this situation (Sowe & Zettsu, 2015). By releasing data related to governance by the ministries, the government gains the trust of its citizens. This gives a clearer picture of the public on the expenditure and policy decisions by increasing trust between the public and the government.

In 2009, the president of the USA Barrack Obama, announced the concept of open data (Orszag, 2009). According to him, it is the right of the people of the country to know what is happening in the government and how the government is investing their money. A lot of countries were not in favour of the idea at the time, and Obama got a lot of objections to this announcement. However, following this, in 2012, the UK government launched its own open data initiatives. After that, a lot more developing

and developed countries have become proactive, and they have embarked on their own open data initiatives. These (ODI) open data initiatives release raw data via various data portals that are disparate and difficult to use, especially by the general public. The data files that have been linked on these data portal do not provide the semantic links among the data sources, hence is difficult to draw the nexuses between data entities. This has created a need for techniques and ontologies that can relate the data with useful semantic links and make it easily accessible over the Internet so that the general public can easily access and use it.

2.1.1 Open Government Initiatives

Open government initiatives are becoming increasingly critical among developed as well as developing countries. The aim of open government initiatives is to provide transparency and wide data reuse. The availability of open data has numerous benefits for the public and other stakeholders. However, the open government initiative data is disparate and published in heterogeneous formats, which makes it challenging to link and reuse the data. In recent years, a number of (OGD) open government data movements have sprung up around the world with two major aims, transparency and data reuse. Some examples are Barrack Obama's open data initiatives in 2009 (Orszag, 2009), open government partnership in 2011 (UK, 2013) and Open Data Charter in 2013 (Washington, 2011). These movements have given rise to open data portals such as data.gov, data.gov.uk, data.gov.gr, open.data.al and data.gov.nz to enable stakeholders and residents to obtain information on any particular ministry of the government. The (OGP) open government portal was announced in 2011, at which point only 45 countries chose to become members. However, since then, more countries have joined the OGP, and the number has gone up to a total of 94. Moreover, statistics show that New Zealand is at the 8th place in the OGP index with 68% open data sets (UK, 2013). These facts

demonstrate the fast-growing rate of data transparency and availability in New Zealand. Figure 2.1 shows the 2014 Global Open Data Index of a number of places, where the grey colour cells show available open data sets whereas the black-coloured cells show data sets not yet open.

The main aim of this index is to track whether the data published is accessible to all stakeholders and to measure the openness of the data at a global level. The availability of open data has several valuable benefits for the general public as well as professionals who use data for making decisions and planning purposes. Encoding (OGD) open government data as (LOD) linked open data would enable a user to browse a data source and then to navigate via the links into other related data sources to get all the relevant data in one place (NZ, 2015). Furthermore, linked open data would also help other decision makers, such as business managers, in essential long-term planning activities because they would get the relevant information easily in one place. In summary, (LOGD) linked open government data will increase government transparency and public awareness of government processes.

Linked Open Data

Linked data is a set of guidelines to link related structured data on the web. The data is represented in triple form (subject, predicate, and object). The subject and predicate in a triple are always (URIs) Uniform Resource Identifiers and the object is either literal, such as a string or number, or another URI. Furthermore, linked data refers to create a link between the data that is from heterogeneous sources and in different formats such as (HTML) Hypertext Markup Language, (CSV) Comma Separated Value, (XML) Extensible Markup language, and (XSL) extensible stylesheet language. It is sustained by more than one organisation in diverse locations, or it specifies heterogeneous systems within a single firm. In this way, linked data refers to the data distributed on the web so that it is machine-readable, its meaning is defined, and it is linked to external data

sets as well as within datasets (Bizer et al., 2011). Linked data is an extension of the web with worldwide data space connecting different companies, films, music, scientific experiments, television, radio programs, medical, drugs, clinical, and many more. There are linked data engines that allow a user to crawl the web data, following the link between the data sources and providing expressive query capabilities. Linked data is based on two fundamental technologies:

- URI : (Uniform Resource Identifiers): are the entities that exist in the real world. It also specifies the addresses for documents and entities on the web. The entities of URIs use the http:// scheme, these URI entities can be looked by dereferencing the URIs over the (HTTP) hypertext transfer protocols.
- HTTP : (Hypertext Transfer Protocol):provides the universal mechanism to retrieve information or resources that is sequential as a stream of bytes (for instance, an image of a flower) or a description of an entity that cannot be sent across the network (for instance, the flower itself)

Method of Linking

There are several possible formats that are used to publish OGD, such as CSV, spreadsheets, HTML tables, and (PDF) Portable Document Format files. An important issue emerging from publishing data using multiple formats is that the users face problems in linking multiple datasets and initiating a data analysis process. For example, assume that a user needs to find the number of schools located in a particular area and the road traffic in that area in the last 2 years so that he can find out which school is located nearby and the rate of traffic during normal and peak hours on those roads which go to that school. To achieve this, the user will need to open two data sets to complete the analysis. In addition, this becomes more complex when performing analyses with more than two data sheets. Therefore, there is a clear need for an infrastructure that

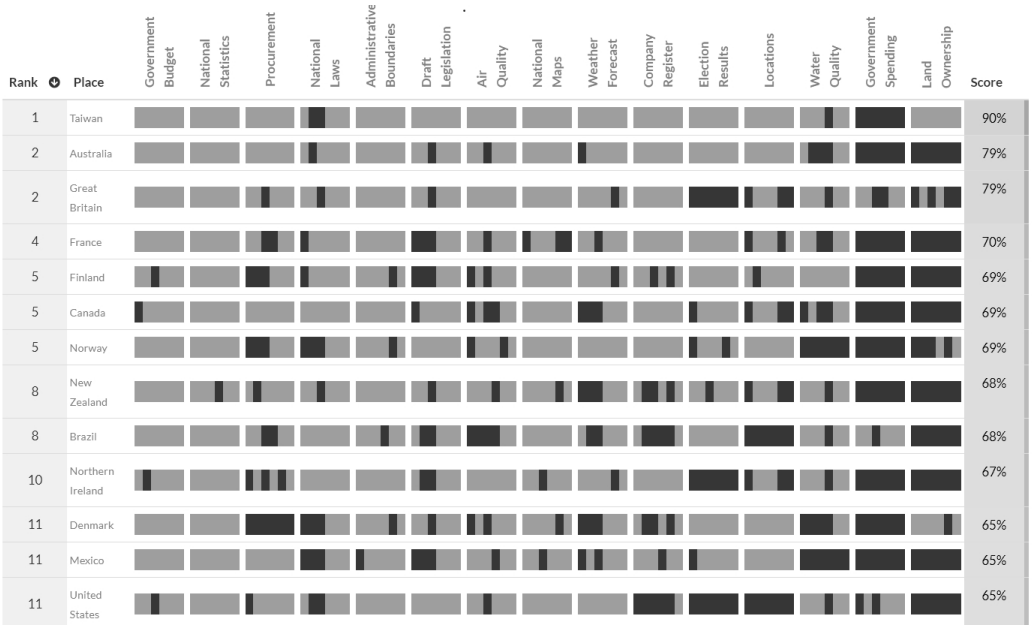


Figure 2.1: Global Open Index (source: https://index.okfn.org/place/) ”public domain”.

links the raw data while preserving the classification, where classification refers to a conceptualization of data according to domains (Heath & Bizer, 2011). One way of data linking and analysis is relational databases, and the other is graph databases.

Relational Database

It is a traditional way of data storage in which tables with multiple rows and columns are created,known as a "record". Each record consists of a set of fields to hold relevant information on a class of objects (Zinke, 2009).For instance, we could have three tables: agriculture, land, and rainfall. These tables could be linked by using fields from one table to another. This would then enable one to extract the information from multiple tables. However, imposing complex queries on these tables is very tedious and requires lots of effort, and if a table has changed, the whole schema needs to be changed.

Table 2.1: Agriculture

Year	Region	Types of Farms
2001	Auckland	Pig
2002	Wellington	Vegetable
2003	Christchurch	Beef

Table 2.2: Land

Year	Region	Types of Land
2001	Auckland	Forest
2002	Wellington	Farming
2003	Christchurch	Volcanic

Table 2.3: Rainfall

Year	Region	Rate of Rainfall
2001	Auckland	2345mm
2002	Wellington	1200mm
2003	Christchurch	2345mm

In the above tables 2.1, 2.2, and 2.3 we have information related to agriculture, land and rainfall. These tables can be joined to access the information via a single query. However, if new rows are added to tables I and III, the query wouldn't be able to retrieve accurate results. Every alteration brings a change in the schema. It requires manual effort to alter the table and perform the join again. Similarly, if any deletion and alternation happens, the whole schema needs to change. It is easy if the tables are small, but very complex, challenging, and time-consuming if the table size increases. Therefore, relational databases have limitation of being static, which can overcome by a dynamic database such as graph databases.

Graph Database

The graph database is an operational database which is designed to operate (CRUD): create, read, update, and delete processes on a graph data model. It uses graph structure for semantic queries with nodes, edges, and properties mainly to store and represent the data (Hitzler, Krotzsch & Rudolph, 2009). It stores the relationship between records. So, a graph consists of nodes and relationships. Each node represents an entity such as a place, a person, thing, category, or other data. A relationship represents how two or more nodes are connected or associated with each other. The graph database uses well defined data models. The most common graph data models are property graphs, hypergraphs, and (RDF) resource description framework triples. Figure 2.2 is an example of a property graph. It illustrates the relationship between agriculture, land, and rainfall in a graph database. The nodes are labelled as agriculture, land, and rainfall. It is connected with a relationship describing how each node is connected. As we see below, vegetables depend on land and rainfall for farming, where farming depends on the rate of rainfall. Moreover, hypergraphs are isomorphic in nature, and hence can be represented as a property graph but not vice versa. Graphs are a means of drawing information via diagrams.

This method of depicting graphs is precisely easy to read for humans if the graphs are small. But, with thousands or millions of nodes, it will be difficult. Moreover, to store and process these graphs in computer systems is also a complex task. Therefore, to extract information from a large graph, we need to divide it into smaller parts where each part can be stored independently. The transformation of such complex data structures into linear strings is known as serialization (Hitzler et al., 2009).

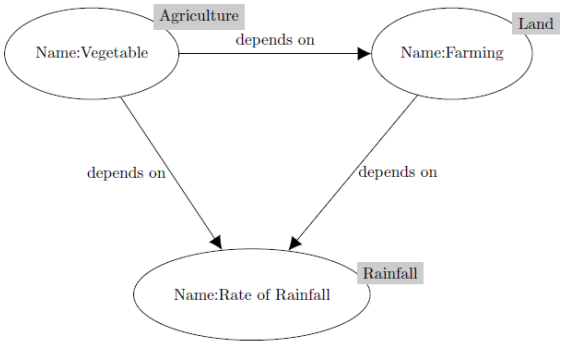


Figure 2.2: An Example of the Property graph

RDF Triples

RDF is a type of graph database which stores semantic facts and information. RDF is a model of data publishing on web standardised by the World Wide Web Consortium (W3C) which supports semantic queries. The main aim of RDF is to allow applications to transfer data on the web without compromising their original meaning. Furthermore, data in RDF is stored in triples, which consists of three elements: subject, predicate, and object. The RDF format is able to take any subject or concept and relate it to any other object using the predicate (verb), which shows the type of relationship between the subject and the object (Barati, Bai & Liu, 2017). An RDF document is a directed graph where both nodes and edges are tagged with identifiers, which are also known as URIs.

Generally, URI's are used for the subject and predicate, where the object can either be another URI or a literal such as a string or number. Further, a literal can have a type that can be a URI. This specifies that triples can have up-to 5 bits of data. Figure 2.3 is an RDF graph with three triples, that specify agriculture depends on land. Further, vegetable and farming are literals of agriculture and land URI's respectively. Agriculture depends on land which specifies that vegetable is a type of agriculture that depends on farming land.

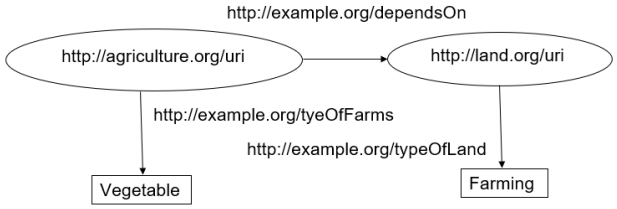


Figure 2.3: An RDF graph describing data values with literals

In addition, RDF triples are written in turtle as follows:

<http://agriculture.org/uri>

<http://example.org/dependsOn>

<http://land.org/uri>.

<http://agriculture.org/uri>

<http://example.org/typeOfFarms>"Vegetable".

<http://land.org/uri>

<http://example.org/typeOfLand>"Farming".

Here URIs are written in angular brackets, quotation marks are used for literals and a full stop is used to terminate all statements.

Semantic Links

An RDF triplestore is used to publish and manage the linked open datasets such as GeoNames and DBPedia. The RDF encoding provides fast query retrieval by using SPARQL (W3C, 2017), which is a semantic query language similar but not the same as (SQL) Structured Query Language. It can be used to retrieve and manipulate the data stored in RDF, which is the basic representation format for the Semantic Web. An ontology supports the organisation of linked data based on the conceptualization. Semantic links bring the concept of linked open data, in which URIs have been created between data files so that different departments' data can be linked together based

on the same fields available in the data files. This will solve the problem of missing links, and the data will be machine readable. More than 70% of the data on the web today is unstructured text (Kucera & Chlapek, 2014), (Ubaldi, 2013). This applies to government data as well, where a large quantity of data is hindered in natural language text documents, making it difficult for humans to understand quickly.

Ontology

An ontology can be coded using both OWL (Web Ontology Language) and RDFS (RDF Schema). However, OWL is more expressive than RDFS, hence it is more suitable for small-scale ontologies (Hassanzadeh, 2011), (Bauer & Kaltenböck, 2011),(Ngomo, Auer, Lehmann & Zaveri, 2014). Ontologies are used to process, capture, reuse, and communication of knowledge. It can be defined as a ”specification of conceptualization”. The domain structure is captured by an ontology, conceptualization represents knowledge about the domain. An ontology is the study of entities in real life and the relationship that these entities have with each other. It is used in a variety of fields, including life science, artificial Intelligence, libraries and, most recently in computer science (Guarino, Oberle & Staab, 2009).

The usage and purpose of ontologies vary from application to application (Taychatanompong & Vatanawood, 2019). The entities in an ontology encapsulate the concepts while the taxonomy represents the relationship between them. In recent times, domain ontologies have become an integral part of numerous knowledgebases and semantic applications (Zong et al., 2015). However, the design process of such ontologies and publishing them with the correct individuals is a tedious and time-consuming task that requires open data as an input so that triples can be generated to make them linked open data. Linked open data and semantic web technologies provide the mechanism for sharing the knowledge that comes from diverse sources (Bizer et al., 2011). According to Tim Berners Li’s principle of five-star linked open data. The combination of RDF,

URI (Uniform Resource Identifier) and SPARQL query makes 5 star linked open data. Figure 2.4 depicts, Berner's 5 star Linked Open Data rule. In nutshell, graph databases have significantly better performance in structured type queries and full-text character searches as compared to relational databases (Vicknair et al., 2010).

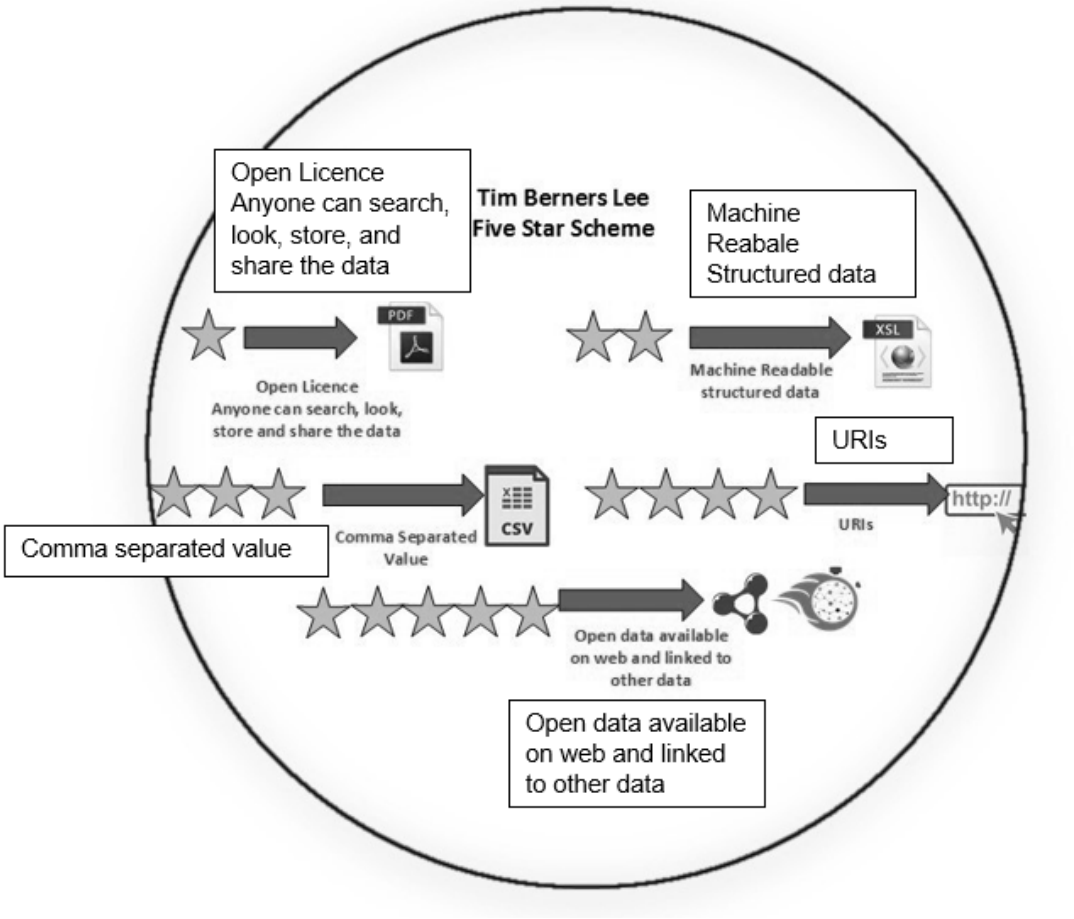


Figure 2.4: 5 Star Linked Open Data (figure redrawn from (https://5stardata.info/en/) (CC BY 1.0 Universal).

2.2 Related Work

2.2.1 Linked Open datasets using String matching

The first systematic study of linked open datasets was reported by (Hassanzadeh & Consens, 2009) which connects several existing movie web resources. String matching measures are used to discover similar links among movie datasets. For evaluation, all string matching results are compared using the owl relationship type, ”sameAs” links. Therefore, the exact matching of movies is not so possible and can give small amount of similarity, hence false results. Additionally, by providing related links about the entities along with the Meta data can enhance the results.

The accuracy of the results is obtained manually by finding all the matching. Thousands of links are inspected manually, which is a time-consuming process. However, if an automatic mining techniques for the data sources are provided, it will save the time. Moreover, more internal links can be provided among the related entities, such as movies with same title. The links can be found using similar string matching techniques. Nevertheless, the linked movie database can be extended further to make it easy to use so that the users can provide feedback on the quality of the links.

2.2.2 LOD Algorithms for Ontology Alignment

A number of researchers have reported linked open data algorithms for the alignment of ontologies of data sources such as, zoology, geospatial, and genetics (Parundekar, Knoblock & Ambite, 2010) and data sources GeoNames, LinkedGeoData, DBPedia, GeoSpecies, and GeneID. In order to determine ontology alignment, an approach has been discussed in which extensions of the classes are compared and restriction classes are defined over the ontology. When the source of the ontology is elementary, restriction classes help to get the redefined set of classes. In an ontology, restriction classes are

used to identify existing and derived sets of classes. Moreover, five pairs of data sources are used to evaluate the alignment of the algorithm by using experimental evaluation.

In this evaluation, linked instances are used but properties not generating useful restricted classes are removed. This provides the reduced data sets, which highlight the applied usage of the equivalence links and properties pertinent to the domain. However, by creating alignment theories, the scalability of this approach can be improved. Therefore, experimental investigation of the space will enhance the performance of the algorithm. This algorithm can be applied to align data resources in biomedical contexts.

2.2.3 LOD Algorithms to detect hidden links in datasets

(Le, Ichise & Le, 2010) has designed an algorithm which is used to detect the hidden links in US Census, DBPedia, GeoNames and World factbook. This algorithm converts linked data into graphs, where nodes represent URIs and edges represent links between URIs. The designed algorithm has been compared with the naive Bayes algorithm, which uses only URIs names to make predictions.

The proposed algorithm resolves the problem of ambiguity in terms of different geographic locations with the same name. In order to get accurate results, multiple patterns of graphs were generated. Missing entities were creating a problem, so a separate pattern was created and the algorithm was applied to predict new entities that can be linked to the missing entities. However, a lot of uncertainty still exists about the usage of the naive Bayes algorithm for comparison with the proposed one. Also, the challenge is to detect hidden links between well-matched URIs and noise.

2.2.4 LOD Prototypes implementation using OGD data sources

Other studies have proposed architectures for indirect determination of links between entities to build linked data. The main aim of this was to link different OGD sources by

creating owl: SameAs links among the URIs. Further, a prototype scenario is designed in which three actors are considered: a school, the local directorate of secondary education of Athens, and the ministry of Education. Five implementation steps are performed to access the consumption of data related to a specific school. SPARQL endpoints are used to access the specific datasets (Kalampokis et al., 2011). However, the whole process is manual, which is time-consuming.

A Silk framework, which is an open-source framework for integrating heterogeneous data sources, has been discussed, but there are no details on how it has been implemented. Further investigation is required to understand the relationship between the data models and political priorities. (Liu et al., 2011) projected a prototype implementation-based case study for linking the Australian government data for sustainability science research. Data sets of energy consumption, population, economics, rainfall, and temperature are used to examine the practical value of linking the data of the Australian government.

The main focus is on reusing published data sets. Numerous challenges were encountered related to data description, discoverability, and analysis. However, the process of discovering relevant datasets is still manual. The automation of this process will enhance the analysis functions of linked data. Moreover, SPARQL endpoints are discussed partially but not implemented for data access. The discussed studies so far have only focused on implementing algorithms and prototypes using linked open data.

The work reported by (Haslhofer & Isaac, 2011) goes a step further and implements the first version of linked open data using open datasets from the libraries, archives, and museums sectors. The technical architecture of the linked open data pilot comprises (ESE) European Semantic Elements metadata. Semantic annotations are created using the Apache Solr tool and all Meta data files are accessible via the Europeana data portal. At the second layer of the architecture, ESE has been converted to the Europeana data Model (EDM) where, dumps are created from semantic enrichment and ESE to EDM conversion.

Furthermore, the dumps are stored as RDF, which can be accessed by the data portal of Europeana. Whenever, the linked data clients access data in RDF specific internet type from data.europeana.eu. The request will be accepted; otherwise, HTTP requests are redirected to the European portal. Using this architecture, one can access the meta data about Europeana resources. However, this alignment requires further evaluation so that it can be extended to other data sources such as DBPedia and other relevant initiatives.

In an analysis of open data for e-government, (S. A. Theocharis & Tsihrintzis, 2013) discussed the Greek open data initiatives' opportunities and benefits. The authors have debated the challenges of opening the data to the public, including its availability, accessibility, reuse, simplicity, global participation, and redistribution. Moreover, the basic steps for converting data to open data have also been highlighted by proposing an architecture for linked open government data.

2.2.5 Internationalization of Linked Data using Framework implementation

The study of linked open data is further carried out by (Kontokostas et al., 2011). They examined the internationalisation of linked data where the language-specific DBPedia framework has been implemented to publish linked data in non-Latin languages. However, it is challenging to create manual SPARQL queries using URIs in non-Latin languages. The aim is to make it easier for the incredible amount of information in DBpedia to be used in new and interesting ways and to inspire new mechanisms for linking and navigating.

The proposed work provides new ways to improve Wikipedia because DBpedia serves as a vital programmatic interface for Wikipedia. Nevertheless, it is challenging to create manual SPARQL queries using URIs in non-Latin languages. It gives rise to

the problem of consistency in terms of triple serialisation formats.

Another example of work using linked open data was conducted by (Alexopoulos, Spiliotopoulou & Charalabidis, 2013). They demonstrated a methodology to analyse the open data movement in the public data landscape from three different perspectives: semantic, functional, and technological. This gives statistical results on the availability of open data in each perspective. However, the government bodies consider their websites to be more advanced than the technical architecture that opens up their data to others to use. If the government focuses on providing a reliable, simple and easy-to-use technical architecture to expose the data, it will encourage the private sector to share useful information or knowledge with stakeholders.

Other studies by (González, Garcia, Cortés & Carpy, 2014) have proposed a theoretical framework to analyse open data based on supply and demand mechanisms. Data sets from the investment portfolio of the Mexican federal budget have been used to develop an example of data visualisation. Tableau software is used to visualise the resultant graphs. These visualisation results will enhance the decision-making process. While accessing the graph, users can apply numerous filters related to the details of a project such as nature, resources allocated, and branch. The graphs are user friendly and help the non-technical users access the information. However, this case is related to Mexico only and how the proposed framework can benefit other stakeholders and countries needs to be explained. Additionally, (Attard, Orlandi, Scerri & Auer, 2015) discussed the opportunities, challenges, issues, and hindrances of open government initiatives. Moreover, guidelines have been discussed to publish the open data. This systematic study also, identifies the impacts on the stakeholders who are involved in the usage of open data initiatives.

In the frequent study (Fragkou, Kritikos & Galiotou, 2016) implemented the methodology described in (Fragkou, Galiotou & Matsakas, 2014). The main goal of this implementation is to create a link between open data initiatives by using an interface

that is based on the E-GIF ontology and the Jena framework. Jena is an open-source semantic web framework for Java. It provides an API for data extraction from RDF. Furthermore, SPARQL endpoints are implemented to query the datasets. However, the main focus is on the workings and characteristics of the Jena framework. It seems the proposed methodology is based on the effectiveness of the tool. Further, the pre designed E-GIF ontology is used for creating RDF models. There is no clue how this E-GIF ontology was created.

Another survey, such as that conducted by (Baiyang & Ruhua, 2016), has shown the value realisation of OGD in information policy, technology, and the economy. The impact factors of value realisation of open government data include information policy, technology, investment, infrastructure, and information system management. However, the OGD area is still lacking in theoretical groundwork and experimental study. Moreover, the study of linked open data was further carried out by (Azevedo, Pinto, Bastos & Parreiras, 2015) in the Geographical information system. This is the project of the Brazilian federal government. Two case studies are used to get a cost-effective decision-making process to minimise the damage from the flood. In the first case study, data from diverse sources is Software. Further, RDF datasets are created from CSV files using the D2RQ platform.

In addition, SPARQL queries are implemented and visualised by using the (GIS) Geographic Information System web application. In the second case study, the designed prototype is used to identify the competency by collecting qualitative data from people having informal group discussions. For this purpose, the DB4Trading web application was built so that users can validate data for semantic repository using their own criteria. Although, this framework seems user-friendly, how the damages of flood will be minimised using this framework is not so clear. Moreover, the datasets can be extended to improve the relevancy. The data visualisation application DB4Trading needs to be discussed in detail.

In another major study, (Fragkou et al., 2014) authors demonstrated the applications of linked open data technologies on the data available in the ERMIS Government Portal for Public Administration. The Jena framework is used for the proposed methodology. The HTML web pages has given as input in the Jena framework from which instances of RDF triples have created and passed to the open link virtuoso server. Further, SPARQL endpoints are provided to access the data.

In addition, the Greek and English versions of the page have passed to E-GIF Ontology and then to the Jena framework. Furthermore, a D2R server with a triple (TDS) database store has been installed and is used to pose the SPARQL queries on the triples. However, the interlinking and identification of data is performed manually. A framework that automate the link discovery process can be used.

A survey of OGD in the Russian Federation by (Koznov et al., 2016) analysed the OGD trends by using the (OECD) Organization for Economic Co-operation and Development analytical framework. The study highlights the progress and implementation of OGD portals in Russia. Numerous eservices have been implemented using OGD. However, more efforts are needed to further redefine the process in order to use OGD in a systematic manner. Recent evidence suggests that linked open data is very useful in linking the data of cities to make them smart cities (Consoli et al., 2017).

(Consoli et al., 2017) proposed a data model which integrates data from heterogeneous sources such as public transportation, road maintenance, municipal waste collection, Geo and urban fault reporting. This prototype linked data portal is available on-line, which helps citizens access information under a free license, and programmers can access the ontology and data via SPARQL queries. However, public transport data can be further aligned with advanced open data standards, which are able to capture more details of traffic and passenger needs. RDF data cubes are also used in this prototype because it focuses on publishing multidimensional data through (SDMX) statistical data and meta data exchange, which is an ISO standard that makes it easier to

exchange meta data and statistical data between different organisations.

A small scale study by (Agrawal et al., 2013) was conducted to construct graphs by analysing the availability of open government data. The findings highlight the sectors such as animal husbandry and agriculture that can be used to create a semantic mesh. Moreover, a basic framework for open data can be built using the findings of the survey.

2.2.6 Data Mapping and Visualization tools and Applications to enhance the consumption of LOD

The analysis shows that (Mutuku & Colaco, 2012) conducted a study on an experiment which brings together subject matter experts in education, transport, water, and local country sectors with open data converters and software developers. The purpose of this experiment is to design an approach to identify OGD applications. The main aim of this study is to find the best practise to increase the consumption of open data via tools and mobile applications. The proposed idea of having an application or tool to access open data that can benefit citizens in spite of their literacy level is unique. However, no technical and implementation details have been shared to achieve this toolkit.

A further example of work that uses RDF triples of open government data was carried out by (Hoxha, Brahaj & Vrandečić, 2011). Authors proposed using the XLWrap wrapper tool to create an RDF triple of the open government data collected from diverse sources. The main objective is to provide transparency via data. Data has been collected from various sources and semantic integration techniques are applied to ease the integration and publication of linked data. However, there is no evidence of the fact that from which particular sector's data sets were used to populate the tool. The objective was to achieve data transparency and data visualisation which has been attained. Therefore, by adding the details of the RDF triple creation and population on the web, the process can become more effective.

By drawing on the concept of the DadosGov catalog, authors (Breitman et al., 2012) have been able to show that the Triplify tool can be used to convert the XML, (JSON) JavaScript Object Notation data sets into RDF triples. Triplify is used for conversion because this tool only takes relational databases as input and produces output as RDF triples. The mapping file defines how database schema concepts must be presented in terms of RDF classes and properties. The creation of data mashups is a time-consuming and complex task. In order to make a comparison between the datasets of Brazil and the USA another set of RDF triples has been created by considering the vocabulary defined in data.gov and the database schema of DataGov. However, the finding shows that the tool doesn't provide necessary support during the conceptual modelling stage.

The study of linked open data was further carried out by (Bahanshal & Al-Khalifa, 2013), who discussed the linking and publishing of linked data clouds and Arabic content on the web. Users can access complex semantic data by querying in Arabic. Currently, there is no DBpedia chapter for Arabic DBpedia, which can act as a source of knowledge. The proposed mapping of Arabic DBPedia with ArabicWikipedia pages will help users access complex semantic data by querying in the Arabic language. Further, natural language tools have been used to extract information from the Arabic DBPedia. However, there are some inconsistencies in design which can be improved by enhancing the knowledge of Arabic DBPedia. Moreover, to extract the English version of the datasets it can be linked with English DBPedia.

2.2.7 Linked open data Ontology implementation of public data

An RDF triple, by definition, consists of a subject, predicate, and object, which can be used to create the classes and properties of an ontology. The first systematic study of linked open data ontology of public spending using triples was reported by (Vafopoulos et al., 2012). It was designed using the data.gov.uk data portal. The input to the

proposed architecture is given by Diavgeia, which is an XML based (API) Application Programming Interface and the first Greek Government Open Data Portal. Output can be seen via SPARQL endpoints. However, the ontology has very basic class and relationship definitions. The data properties, objects, concept-restrictions, and rules have not been created for the proposed ontology.

In another study, Theochairs discussed an ontology (S. Theocharis & Tsihrintzis, 2014) using protege 4.2 to link the public administration data. The built-in reasoner of protege is used to find out semantic errors in the designed ontology. This study is an attempt to present a part of the ontology concerning the characteristics of administrative acts. Therefore, the proposed ontology can be extended further by adding more concepts and their properties. It will contribute to forming a knowledge base for the management and development of open data. Furthermore, human evaluation is conducted in the form of posing questions to the ontology using SPARQL endpoints. However, human evaluation can be challenging because, (we) human are disposed to make mistakes.

Another example is an attempt to enrich the Greek e-GIF ontology (Galiotou & Fragkou, 2013) where protege is used to design the entities. A number of entities are added to the existing ontology and a comparison has been made with (PSCs) Point of Single Contact of other European countries such as Cyprus, Malta, Spain, and the Slovak Republic. However, more attempts are required for further enrichment of the entities. A refined version of ontology along with good comparison results will be used for semantic enrichment of higher elements of the web pages with URI properties. This is vital for the conversion of open government data into linked data.

Some further examples of linked open data are exploratory study of Brazilian initiatives based on the principles of linked open data is conducted by (Matheus, Ribeiro & Vaz, 2012). Brazilian portals are three-star, which means data sets are in XML, CSV, and HTML. In addition, (Hendler, Holm, Musialek & Thomas, 2012) has done discussion on the data.gov portal and the use of linked open data. They have focused on

various sectors where linked data is utilised.

In an analysis of linked open data, (Zhao & Ichise, 2014) retrieved graphs based ontology from the diverse data sources available publicly. The ontology alignment methods have been applied to identify classes and properties of ontology from the data sets. Related classes and properties from different datasets are combined to find the missing ”SameAs” links. This semi-automatically created cohesive approach solves the heterogeneity problem of ontology. Moreover, it finds the missing and wrong properties in the data sets. However, only four datasets have been selected, which are DBPedia, GeoNames, NYTimes, and LinkedMDB. In order to extend the alignment process, more data sets are required, and the MapReduce method can be used to deal with big data sets.

In an attempt to semantify open data, (Al-Khalifa, 2013) proposed a lightweight approach for re-using existing ontologies from (Hoxha et al., 2011). The main objectives were to contribute to the knowledge of the semantic web and enable data exchange and linking with other semantic sources over the web. However, the conversion process was not fully automated.

Furthermore, there is no open data portal available in Saudi, which gives rise to the problem of data extraction. If common vocabularies to access the data are developed, it will help to link open data initiatives world widely. Moroever, the OGDL4M ontology will be used for qualifying datasets in order to improve the accuracy of their legal annotation. The Ontology also aims to connect each applicable legal rules to official legal texts in order to direct legal experts and reusers to primary sources.

2.2.8 Challenges of Open data

In a study, (Yang, Lo, Wang & Shiang, 2013) demonstrated the opportunities, challenges, and negative impacts of open government data. The six dimensions of the open data

policy of Taiwan have been discussed. The first of the six policy dimensions is the need for Taiwan to have a strong open data legislation policy. Second, current data sets should be extended further. Third, cloud computing should be used as a fundamental facility for the distribution of open government data. Four, a single portal to access all government open data. Fifth, public awareness is a must to utilise open data, and lastly, open data should be free to the public. There should not be any licence fee for downloading the data.

2.2.9 Open data Management Tools and Activities

In another study, (Bojārs & Liepiņš, 2014) discussed (PDH) peak data hackathons, which is an activity initiated by a group of volunteers. The main purpose of PDH is to use available open data and transform that data into user-friendly form. After completion, the data has been submitted to data.opendata.lv. However, the work-done by the volunteers is not static because they have no right to hold and maintain the data. The datasets used do not come under an open licence and are not up-to-date. Therefore, an automation process is required to update the catalogue based on meta data related to open data.

In a case study focused on evaluation, effectiveness, and capabilities of open data of the Canadian municipal level, (Gruzd & Roy, 2014) Roy has analysed open government data and open data governance in Canada, with a particular focus on municipal governments. This will bring systematic transparency and overcome the challenges of conceptualization, the role of the media, political and data culture, and a holistic governance framework. However, open and innovative governance is required to bring progression in the open data movement. One of the most significant current discussions is that ODI can be managed using online APIs, tools, and websites.

A survey conducted by (Corrêa, Corrêa & da Silva, 2014) over 20 municipalities

to analyse the access to information law in Brazil. The findings suggest that 10% of the documents are in open format. The usage and importance of the Comprehensive Knowledge Archive Network (CKAN)(International, April,2017), which is used to manage and publish open government data, is also highlighted. A number of research institutions, local as well as national governments are using CKAN. The published data can be previewed in the form of graphs, tables, and maps. It is an open-source data portal platform, so it is used by a number of countries. Internal modes of CKAN have been used to store the Meta data about diverse sources, and they have been presented on the web so that users can search the data. However, the challenges of culture change still have to be dispersed across local governments. Table 2.4 provides the complete information related to the work done in linked open data.

Further, a study of open government data (Mockus & Palmirani, 2017) discloses that legal requirements which are unspecified make open data complex to use. The perplexity of legislative requirements indicates the necessity of having an ontology that facilitates the analysis of the licenses, conditions, and legal notices of OGD. These analysis techniques can be automatic or semiautomatic.

The work (Jiang, Hagelien, Natvig & Li, 2019) proposes a prototype search engine for automatic and manual linking of the concepts in the transport domain ontology. The result has shown quality search and more efficient open data search. However, the proposed prototype can be improved by adding more datasets as well as by adding accurate metadata descriptions.

More recent work by (Escobar, Roldán-García, Peral, Candela & Garcia-Nieto, 2020) used the water supply management datasets of Valencia (Spain) to generate an ontology-based framework for publishing the linked open data. The proposed ontology is capable of identifying the correlations among the water supply, leakage, and population. However, the interlinking process can be improved by adding more data sets of different domains.

Table 2.4: Provides the complete information related to the work done in linked open data

Year	Author	Datasets Used	Evaluation Method	Techniques
2009	Hassanzadeh et al.	Freebase, OMDB, DBPedia movies, Rotten tomatoes.com, Stanford movie database	String matching and precion recall	Weighted Jaccard, Jaccard, Edit Similarity BM25, HMM (Hidden Markov Model), Cosine w/tf-idf
2010	Parundekar et al.	LINKEDGEODATA, GEONAMES, DBPEDIA, GEOSPECIES, MGI, GENEID	Empricial Evalaution	Alignment algorithm
2010	Thanh Le et al.	US Census, GeoNames, DBPedia, World Factbook	Experimental Evaluation	Graph Mining techniques is used in the Algorithm to detect hidden relationship by comparing Naïve algorithm

Continued on next page

Table 2.4 – continued from previous page

Year	Author	Datasets Used	Evaluation Method	Technique
2011	Kalampkois et al.	Moraitis School, The 2nd local Directorate of Secondary Education of Athens, The ministry of Education	Prototype implementation	Silk framework, D2R server, SPARQL interface, RESTful APIs
2011	Liu et al.	Energy, Population, Economic, Environment (Rainfall and temperature data)	Prototype based Case Study	RPI data conversion tool, TWC LOGD portal, cell based conversion
2011	Haslhofer et al.	Libraries, Archives, Museums	Europeana LOD Prototype	Open Link Virtuoso, Dereferencing HTTP URIs, RDF Mapping, SPARQL Queries
2011	Kontokosras et al.	Greek Dbpedia	Dbpedia Informtion Extraction Framework	Infobox mapping and properties, Inter language link extractor, Inter Dbpedia linking, IRI Serialization, Transparent Content negotiation rule

Continued on next page

Table 2.4 – continued from previous page

Year	Author	Datasets Used	Evaluation Method	Technique
2012	Hoxha et al.	OGD of Albania	ODA ontology	XLWrap Wrapper, Ajax, SPARQL, Google Visualization API, Spark (Javascript library)
2012	Breitman et al.	OGD of Brazil related to Population	DataMashups	Triplify, stdTrip
2012	Vafopoulos et al.	Public spending datasets	Ontology	Virtuoso Jena Provider, Python libraries, Diavgia API, Java libraries, XML static instances
2013	Galiotou et al.	ERMIS Government portal for Public Administration	Case Study	Jena Framework, Open linkVituoso, SPARQL Endpoints
2014	Fragkou et al.	ERMIS Government portal for Public Administration	Enrichment of e-GIF ontology	Protégé
2014	Theocharis et al.	Public Administration Greek	Ontology development	Protégé

Continued on next page

Table 2.4 – continued from previous page

Year	Author	Datasets Used	Evaluation Method	Technique
2014	Zhao et al.	Dbpedia, Geonames, NYTimes, Linked-MDB	Graph based ontology Analysis	SameAs Graphs extraction Algorithm, Related classes and properties grouping, Aggregation of all integrated classes and properties
2014	Gonzalez et al.	Mexican federal budget, investment Portfolio	Theoretical framework for Data Visualization	Tableau software Public version 8.1
2016	Galiotou et al.	Agriculture, Wholesale and retail, Transportation, Accommodation and food services, Education, Arts, entertainment and recreation, Construction, Mining and Quarrying	interface implementation using Jena	Jena Framework, FUSEKI, SPARQL Endpoints

Continued on next page

Table 2.4 – continued from previous page

Year	Author	Datasets Used	Evaluation Method	Technique
2016	Azevedo et al.	ANA, ANEEL, IGAM, CPRM, CEMIG, Transparency portal for MG, IBGE, Health Portal, PNUD, Open data Portal, Geonames, DBPedia, Generic data sources (such as Google, Wikipedia, sciencedirect etc.)	case study-based experiment	D2RQ platform, GIS for Visualization of Map and SPARQL Queries, DB4Trading
2017	Martynas Mockus, Monica Palmirani	European Union (EU) legal framework of reuse of Public Sector Information (PSI), the EU Database Directive and copyright framework and other legal sources (e.g., licenses, legal notices, terms of use)	Open Government Data Licenses Framework for a Mash-up Model	MeLOn, RDF, SPARQL Queries, OWL reasoner

Continued on next page

Table 2.4 – continued from previous page

Year	Author	Datasets Used	Evaluation Method	Technique
2019	Shanshan Jiang, Thomas F. Hagelien, Marit Natvig,Jingyue Li	Transport domain datasets	A prototype search engine for manual and automatic linking of the concepts in the ontology	Semantic, NLP (Natural Langauge Processing), Semantic and Machine learning techniques
2020	Pilar Escobar, Maria del Mar Roldan-Garcia, Jesus Peral, Gustavo Candela and Jose Garcia-Nieto	Water Supply management of Valencia (Spain)	An Ontology Framework to generate semantically enriched linked data	RDF, OWL, SPARQL, Data Mapping, Preprosesing, Data Modelling, storage and Exploitation.

2.3 Benefits of Linked Open Data

2.3.1 Transparency

Linked open data brings transparency among the various sectors of the government. In addition, ordinary citizens of the country would be able to see the performance of the various government departments. Every sector's data will be accessible at one place which will help the users to navigate amid the links to extract the useful information (Walsham, 2001). The citizens will be able to participate in political and social activities because, once the information is readable and understandable, it will encourage the citizens to show more interest in political and other social affairs.

The linked open data will create trust among the general public towards the government (Attard et al., 2015). Everyone will have equal access to information, which will help the government improve services and processes for citizens. It means all stakeholders can see, use, and distribute the data freely. In order to achieve a transparent democracy, the country should open up the data so that the stakeholders can see the activities of the government (Zuiderwijk & Janssen, 2014).

2.3.2 Public Participation in Government

Open data provides an opportunity for the public to participate widely in government activities. The general public can impose questions on certain areas. It will assist the government in making more equitable policies and making better decisions (Bertot, Jaeger & Grimes, 2010). For instance, if the government wants to introduce a new change in current policies and intends to seek the viewpoints of the public by disclosing the operational cost and accompanying benefits. This will help the people to make a clear and firm decision in terms of supporting or opposing the decision (Janssen et al., 2012).

2.3.3 Social Value

The government is responsible for handling a large amount of data relating to education, health, the environment, agriculture, employment, and budgets. By opening this data to the public, the government can encourage people to use it and invent new ideas, thus helping others create social value (Zuiderwijk, Janssen, Choenni, Meijer & Alibaks, 2012).

2.3.4 Reliability

Open data initiative will lay down the foundation of a strong and open government which work and care for people. The data available by the government will help people to believe in the government. They can take clear decisions based on the data sources provided by the government (Saxena & Muhammad, 2018).

2.3.5 Economic Growth

One of the biggest benefits of linked open data will be economic growth and the stimulation of competitiveness. It will stimulate innovation and contribute towards the improvement of various processes, services and products. Moreover, the information will be available for various investors and companies, which will help them to invest their money wisely, and ultimately, it will add value to the economy (Kalampokis, Tambouris & Tarabanis, 2013).

2.3.6 Data Sustainability and Reusability

This is the most generic criteria for the quality of open data. Reusability means how easily the published data can be reused. It means the data sets collected once can be used multiple times without recollecting the data. This will reduce unnecessary duplication

and associated costs. In addition, open data will bring sustainability to data (Zuiderwijk & Janssen, 2014).

2.3.7 Decision Making

Open data will have a huge impact on fair decision making because it will allow people to compare different options.associated data sources. The new data sources can be created by combining the available or existing data sources (Kucera & Chlapek, 2014).

2.3.8 Integration and Availability of Information

Open data will bring the concept of information availability. All vendors and companies can access the useful information from the data portals. The operational and technical benefit of open data is to integrate, merge and mesh private and public data (Kucera, Chlapek, Klímek & Necaskỳ, 2015).

2.4 Challenges of Linked Open Data

2.4.1 Data Protection and Privacy

There is a conflict between the aims of transparency, accountability, open data, and data protection. Even though the data is nondescript before merging and publishing, there is still possibility that it could result in the discovery of some personal information. For example, if garbage collection data is published alongside a personnel timetable, the data consumer will be able to easily identify the route of a specific employee. Therefore, this issue requires more research in order to come up with some predefined guidelines that can provide a solution to this privacy concern. Furthermore, imposing some restrictions on data access could solve this problem (Palmirani, Martoni & Girardi,

2014). However, applying this approach will restrict the openness of the data on several levels.

2.4.2 Complexity

Open government initiatives are growing so rapidly, but there is still a lack of ability to discover the appropriate data. Data consumers can only access the processed data. Moreover, the meaning and explanation of the data is highly complex. Users have difficulty in browsing and searching due to lack of an index or other useful means to ensure easy search for finding the right data. The data format and datasets are too complex to read and use. There is no support to extract information from such big initiatives. There is a focus on one dataset. Therefore, the actual benefit will come from linking various datasets together (Bauer & Kaltenböck, 2011).

2.4.3 Lack of knowledge

The users of open data have no motivation to access the information. There is a lack of knowledge among the citizens to make use of the data. They are lacking the capability to use the information. Furthermore, they are not interested in the data for a variety of reasons, including licence fees, big data initiatives, redundant registrations prior to download, lack of statistical knowledge, unexpectedly increased costs, unsupported behaviour of public organizations, and threats of lawsuits (Janssen et al., 2012).

2.4.4 Legislation

Although most of the open government initiatives fit into existing legal frameworks, there is a lack of an open government policy. he problem is that security and privacy are being violated (Dawes & Helbig, 2010). Most of countries do not have any licence on the use and consumption of open data, while others prefer to have written permission

required prior to gaining access to the data. In this way, an authenticated person will get access. The written agreement between the data consumers and legal entities will provide security. Moreover, if any person breaches the agreement, they will be liable for punishment (Arcelus, 2012).

2.4.5 Quality of the Information

The quality of the information available in the data initiatives is not appropriate. Most of the data is invalid and obsolete. The incomplete information is a challenge to tackle when part of the information is visible and the rest of the information is missing or not loaded (Zuiderwijk & Janssen, 2015). However, the information may appear irrelevant when viewed in isolation, but when linked and analysed, it can result in new insights. Moreover, similar information is stored in several places, which causes confusion among the data consumers.

2.4.6 Technical

Open data initiatives are facing plenty of technical challenges. It is problematic to access the data if it is not readable by the machine and the format is not well defined. Currently, there is no architecture and standards for the processing of open data (Yang & Wu, 2016).

2.5 Motivation

The tabular data from the real world is revealed over the web to provide an open platform for linking to other data resources. As a result, this information is stored in an obsolete data format. Due to its complexity, it is difficult to import data into a web context in order to conduct reasoning and analysis on it. In order to carry out reasoning and

analysis, the machine must be able to comprehend the semantic details of the data. Semantic links bring the concept of linked open data, in which URIs have been created between data files so that different departments' data can be linked together based on the same fields available in the data files. This will solve the problem of missing links, and the data will be machine readable. More than 70% of the data on the web today is unstructured text (Kucera & Chlapek, 2014), (Ubaldi, 2013). This applies to government data as well, where a large quantity of data is hindered in natural language text documents, making it difficult for humans to understand quickly.

For example, if a new entrepreneur wants to buy a business and he is not sure exactly how to select an appropriate business that will bring more profit to him, LOD could be helpful in making the decision to select the right business. To do this, data from the business and financial sectors would be linked together using LOD principles. This will show how the business is progressing and how the year went in terms of the business. Thus, end users would be able to make worthwhile decisions. Moreover, LOD will provide transparency to the general public, so that individuals can get to know what is happening in the country.

For instance, by making the annual financial spending data available as LOD, the public will have access to the data and they can easily figure out where the money is being invested. As a result, government agencies must justify their operations by providing data sets that provide a clear picture of all spending and reduce the likelihood of bankruptcy and embezzlement. Hence, there is a need to have an approach that can assist in the interlinking of diverse datasets where semantic encoding can be achieved.

2.6 The Proposed Design Architecture

The Figure 2.5 depicts the proposed architectural design. The whole process is divided into three phases:

- CSV to OWL Conversion
- Generating the Semantic link
- Implementation of the SPARQL interface

The CSV to OWL conversion process starts when the user enters either the URL of the CSV file or uploads the CSV file directly. The user has to mention the name of the CSV file. Later this name will be used as the name of the ontology. The subsequent concepts associated with each CSV file are created. To read the CSV files effectively, the Apache Commons CSV library is used, as it defines the format for the CSV file. Here, default mode is used to parse the CSV files so that all headers and records in the CSV files can be parsed successfully.

Once the CSV file is parsed, interpreted CSV tables are generated, the CSV to OWL converter process is initiated, which helps in transforming the CSV datasets into OWL triples. These OWL triples are stored in the local memory of the system, which can be visualised using the protege tool. To generate the semantic link, the user must choose two or more ontology files. Because ontology files can contain many ontology classes, it is necessary to select the appropriate class from the ontology file.

Additionally, the user must select common data properties from both ontology classes in order to perform semantic link operations. Once the user has generated all of the aforementioned ontologies, the user can choose between two distinct ontologies and create semantic links between them. Following the selection of two ontology files, the files must be combined to produce a union. In this case, the union contains all ontology classes, individuals, and data properties.

The primary goal of generating the union file is to retain all the information contained in the ontology files in order to avoid data loss. The class of the second ontology file is converted to tabular form as a step of the procedure by employing the HSQL Database. The purpose of this conversion is to simplify the search operation of the semantic links

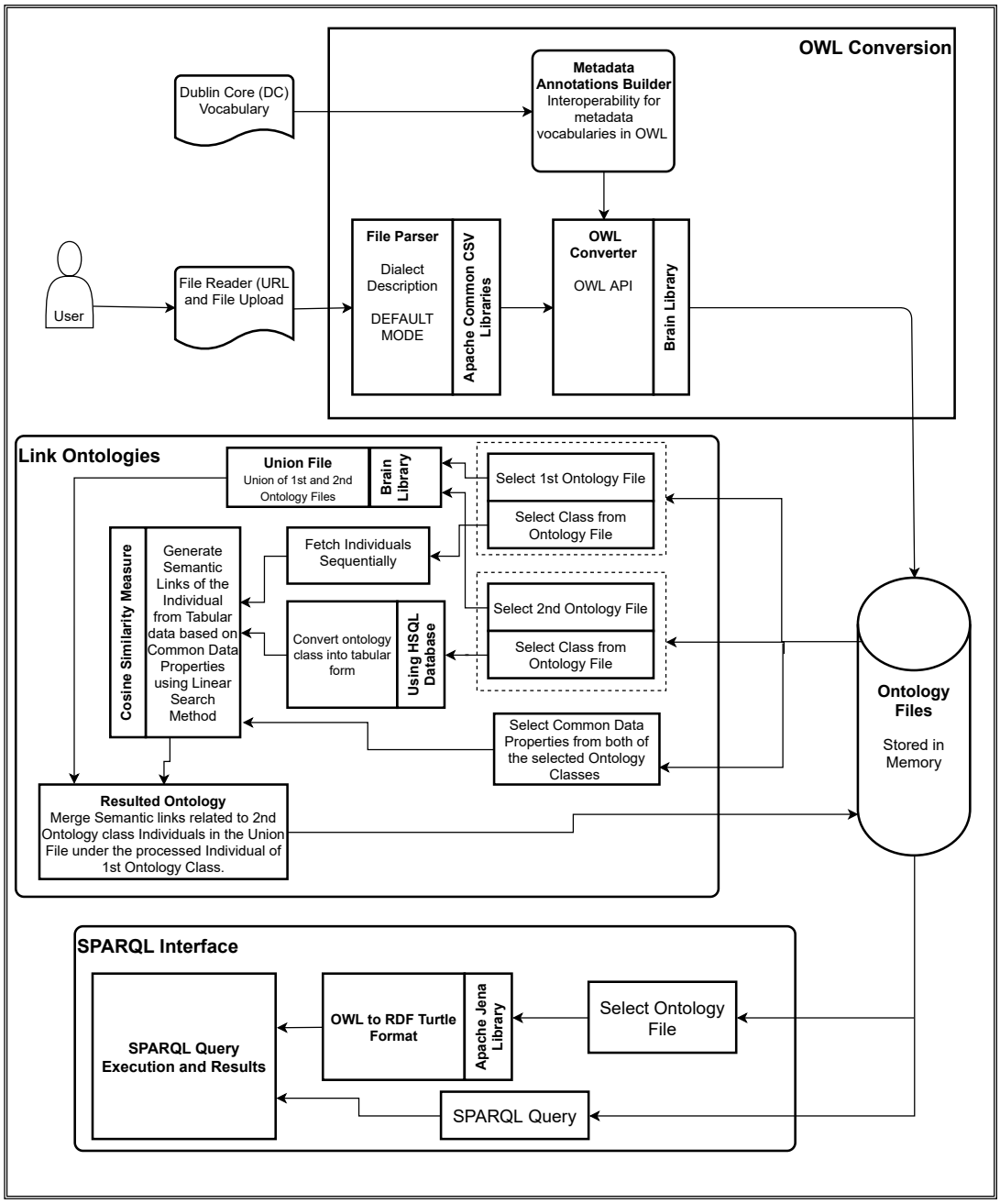


Figure 2.5: The Proposed Architecture

by utilising the tabular form, which is simple to traverse when performing a linear search. The following step is to iteratively or sequentially retrieve the individuals from the first ontology class.

Additionally, the data properties of the first ontology class individuals are compared with the tabular data by performing a linear search operation so that semantically linked individuals of the second ontology class can be identified. Individuals may have duplicate values, making it difficult to semantically link these datasets based on common properties.

As a result, an appropriate approach for linking these ontology classes is based on a condition that determines the literal values of data properties of ontology classes that are similar. We overcame this issue by utilising the cosine similarity measure, which includes methods for comparing two strings and returning the similarity score. The linear search method locates individuals inside tabular data in order to establish semantic links between selected common properties. It will continue to browse through each piece of data progressively until a match is discovered or the full table is searched. The produced semantic links are then integrated with the union file's second ontology class individuals. As a result, a semantically linked ontology for two distinct domains or data sectors will be created. Moreover, the semantic linking process is one-directional only.

In future versions, bi-directional linking will be considered. Once we have constructed our semantically linked ontologies, the third and last phase is to use the SPARQL interface to query the produced ontology. The SPARQL interface is structured in such a way that the user can choose the ontology file to query. To query in SPARQL, the created OWL ontologies must be converted to an RDF format, such as Turtle. Composing SPARQL queries that contain complex OWL expressions varies in complexity from difficult to inconvenient due to the fact that SPARQL query syntax is built on Turtle (D.Beckett, 2020), which is not designed for OWL. We used Apache Jena for the

conversion process; it's a java package that converts OWL files to RDF Turtle format and offers APIs for querying SPARQL from within a Java application. A diagram of the SPARQL interface is illustrated in Figure 2.5.

2.7 Case Study

The approach described in Section VII has been evaluated using the data sets for agriculture, land and rainfall from the New Zealand government. The aim of our work is to introduce a prototype which will assist in automating the process of ontology creation by using government datasets. The datasets are extracted from the URL https://data.mfe.govt.nz. The file parser parsed the entered URL, and the CSV datasets are converted to OWL files using an OWL converter. All OWL files are saved locally in the system memory. A visual representation of agriculture ontology can be seen in figure 2.6. Owl:Thing is the main/default class here, and Agriculture is the subclass of it, which is holding five data properties such as area_ha, farm type, region, FID, year, and 620 individuals. When the mouse pointer is hovered over an individual, the data property assertions for that individual are highlighted.

Likewise, an ontology is developed and stored for land and rainfall datasets. Furthermore, these generated ontologies will be semantically linked so that knowledgeable data may be retrieved by implementing SPARQL queries. To perform this task, two ontology files from the local hard drive of the system are selected, and then the shared properties are manually identified. The common data properties in both ontologies are year and region. As a result, these data properties are chosen for the generation of semantic links.

Furthermore, we do not want the ontology files to lose any information, so a union file is formed by joining both of the ontology files. This union file was set aside and eventually combined with the resulting ontology. Here we have selected the agriculture

ontology as the first ontology file and land as the second ontology file. If the ontology has more than one class, we must ensure that we select an appropriate class from the ontology files. We selected the agriculture and land classes from the agriculture and land OWL files, respectively. Individuals from the agricultural class are retrieved one at a time, whereas the land class has been converted into a tabular form to simplify the semantic link search operation because the tabular form is simpler to traverse while executing a linear search operation.

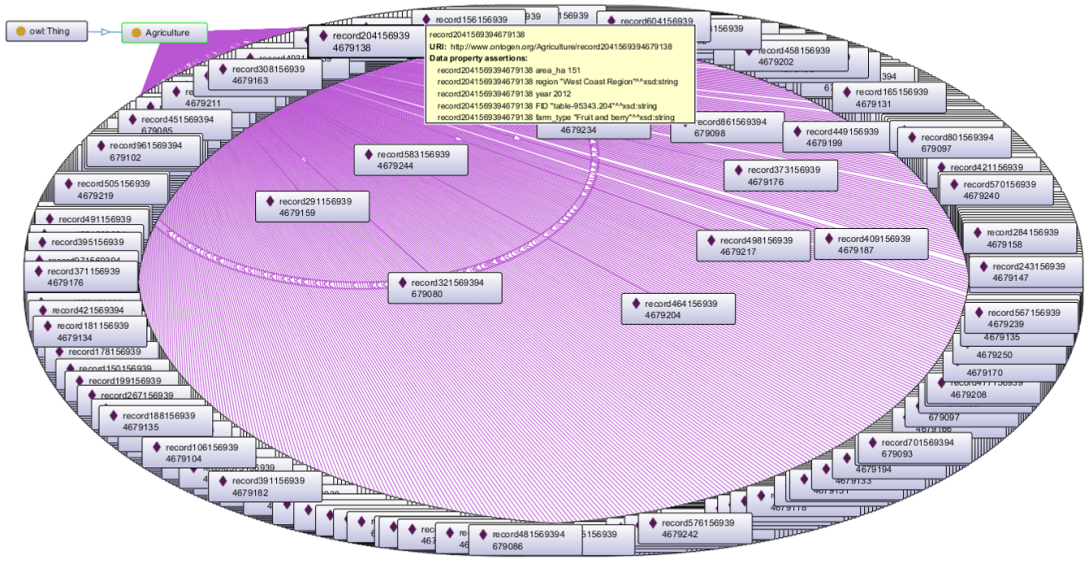


Figure 2.6: The Agriculture Ontology

inally, the semantic links are constructed by conducting a linear search on individuals in the agricultural class and tabular data from the land class. To avoid duplication, cosine similarity is employed. Prior to getting the full version of the linked ontology, the previously formed union file is merged with the resulting ontology file, allowing us to have both semantic and non-semantic data together. Figure 2.7 shows a photographic image of the semantically linked ontology of agriculture and land datasets.

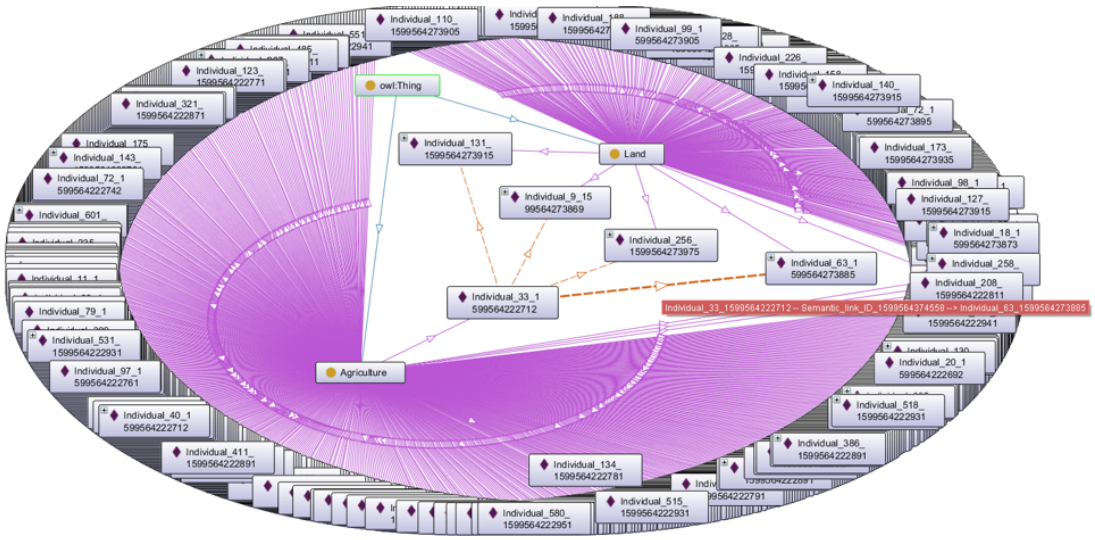


Figure 2.7: The Semantically Linked Ontology of Agriculture and Land

The dashed arrow lines represent the semantic link among both the individuals of the land and agriculture classes. There are 892 individuals in total, with 168 semantic links. Due to screen size constraints, we are unable to display all of the semantically linked individuals.

To evaluate our methodology, we must first examine the accuracy of the created ontology. We did this by feeding the framework certain sample SPARQL queries. The results of certain queries are recorded, and the consistency of those results is carefully assessed. SPARQL queries are performed on datasets to collect relevant information. Traditional data extraction approaches are time-consuming as they require a comprehensive reading of the datasets. However, as all data is stored as triples, an RDF query language makes the process simple. It is straightforward to find accurate data by using SPARQL queries. For testing purposes, we applied the following SPARQL queries to the resulting ontologies:

1. Find the rainfall rate of wellington region in 2007
and what type of farming activities are carried out in
that year?

```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-
ns#>
SELECT DISTINCT ?ind1 ?ind2 ?Rainfall ?farm_type
WHERE
{
?ind1 rdf:type owl:NamedIndividual .
?ind2 rdf:type owl:NamedIndividual .
?ind1 ?linked ?ind2 .
?ind1 rdf:type <http://www.ontogen.org/Rainfall/Rainfall>.
?ind1 <http://www.ontogen.org/Rainfall/Year>2007 .
?ind2 rdf:type <http://www.ontogen.org/Agriculture/
Agriculture>.
?ind2 <http://www.ontogen.org/Agriculture/year>2007 .
?ind2 <http://www.ontogen.org/Agriculture/region>
"Wellington Region".
?ind1 <http://www.ontogen.org/Rainfall/r95ptot>?Rainfall.
?ind2 <http://www.ontogen.org/Agriculture/farm_type>
?farm_type
}
```

2. Find the region with maximum rainfall rate (r95ptot) in year 2012. Also find the farming activities for that region

```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-
ns#>
SELECT ?ind1 ?ind2 ?year ?site ?farming ?rainfall_rate
```

```
WHERE
{
?ind1 rdf:type owl:NamedIndividual .
?ind2 rdf:type owl:NamedIndividual .
?ind1 ?linked ?ind2 .
?ind1 rdf:type <http://www.ontogen.org/Agriculture/
Agriculture>.
?ind2 rdf:type <http://www.ontogen.org/Rainfall/Rainfall>.
?ind2 <http://www.ontogen.org/Rainfall/site>?site .
?ind2 <http://www.ontogen.org/Rainfall/r95ptot>
?rainfall_rate .
?ind1 <http://www.ontogen.org/Agriculture/farm_type>
?farming.
FILTER (?rainfall_rate =?maximumrainfall) .
?ind2 <http://www.ontogen.org/Rainfall/Year>?year .
{
SELECT (MAX(?rainfall) AS ?maximumrainfall)
WHERE
{
?ind3 rdf:type <http://www.ontogen.org/Rainfall/Rainfall>.
?ind3 <http://www.ontogen.org/Rainfall/Year>?year.
?ind3 <http://www.ontogen.org/Rainfall/r95ptot>?rainfall.
FILTER(?rainfall != "NA" && ?year =2012)
}
}
}
```

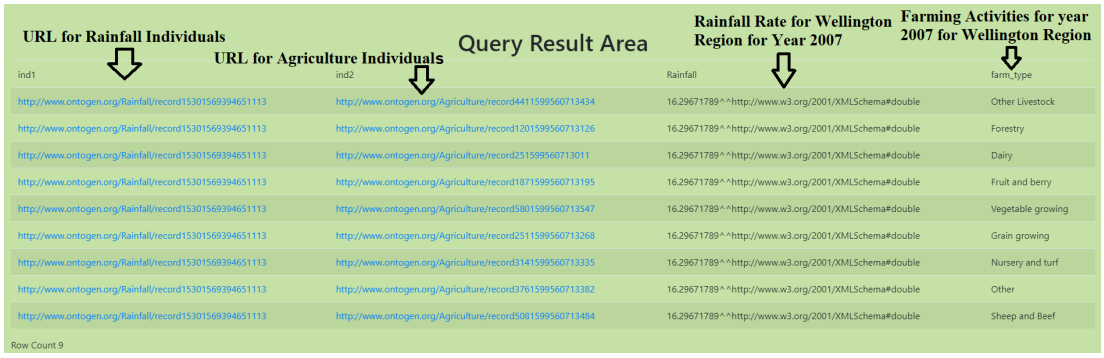
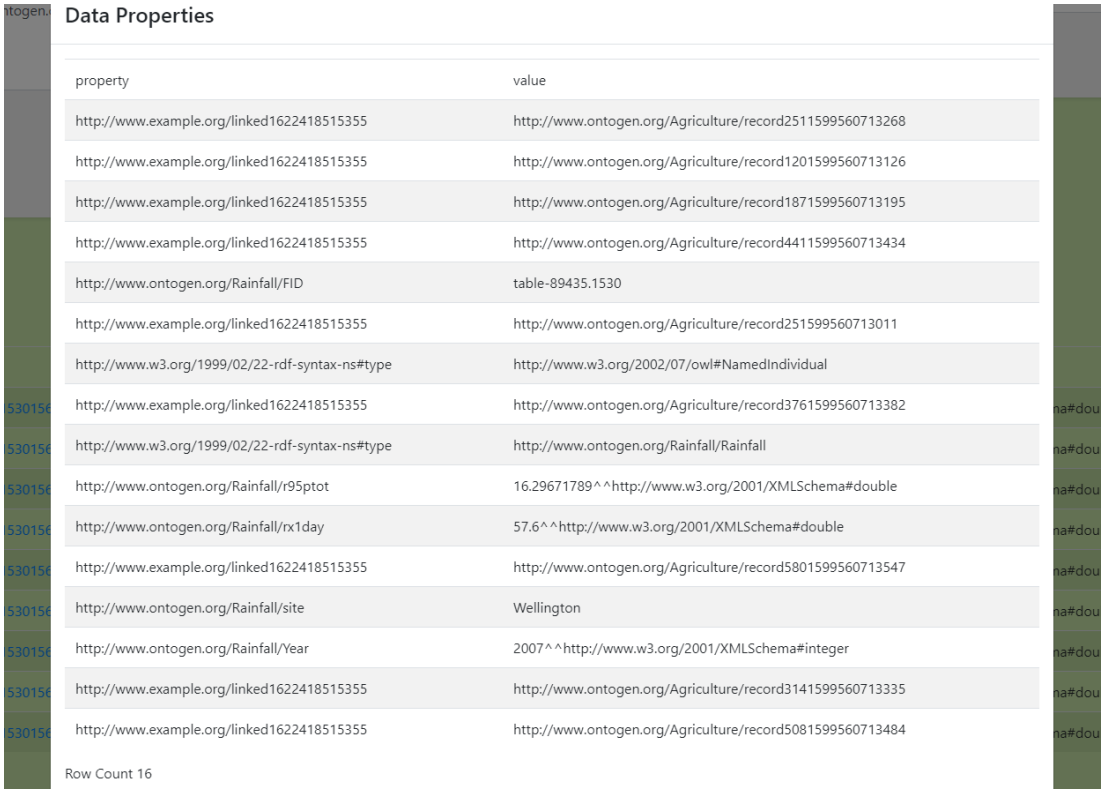


Figure 2.8: Result of 1st Query for Rainfall and Agriculture Ontology

Figure 2.9: Result of 2nd Query for Rainfall and Agriculture Ontology

Figure 2.8 and 2.9 show the results for the queries one and two, in which semantically linked ontologies of rainfall and agriculture are used to impose the SPARQL queries and extract useful information. A manual analysis is undertaken for accuracy and consistency. Following a manual review, it was discovered that the system's results were accurate. As a result, our SPARQL interface interacts with ontologies that are semantically interconnected and capture the desired results. At this point, we're using SPARQL queries to evaluate our results. We want to perform an expert evaluation of the proposed model in the future to ensure its performance and robustness. The individuals highlighted in Figure 2.8 can be accessed by clicking on the blue hyperlink. To access the data properties, one needs to click on the individuals either under ind1 or ind2, the corresponding data properties will be highlighted. As Figure 2.10 highlights all data properties corresponding to the 3rd individual under ind1 of Figure 2.8.



Data Properties

property	value
http://www.example.org/linked1622418515355	http://www.ontogen.org/Agriculture/record2511599560713268
http://www.example.org/linked1622418515355	http://www.ontogen.org/Agriculture/record1201599560713126
http://www.example.org/linked1622418515355	http://www.ontogen.org/Agriculture/record1871599560713195
http://www.example.org/linked1622418515355	http://www.ontogen.org/Agriculture/record4411599560713434
http://www.ontogen.org/Rainfall/FID	table-89435.1530
http://www.example.org/linked1622418515355	http://www.ontogen.org/Agriculture/record251599560713011
http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.w3.org/2002/07/owl#NamedIndividual
http://www.example.org/linked1622418515355	http://www.ontogen.org/Agriculture/record3761599560713382
http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.ontogen.org/Rainfall/Rainfall
http://www.ontogen.org/Rainfall/r95ptot	16.29671789^^http://www.w3.org/2001/XMLSchema#double
http://www.ontogen.org/Rainfall/rx1day	57.6^^http://www.w3.org/2001/XMLSchema#double
http://www.example.org/linked1622418515355	http://www.ontogen.org/Agriculture/record5801599560713547
http://www.ontogen.org/Rainfall/site	Wellington
http://www.ontogen.org/Rainfall/Year	2007^^http://www.w3.org/2001/XMLSchema#integer
http://www.example.org/linked1622418515355	http://www.ontogen.org/Agriculture/record3141599560713335
http://www.example.org/linked1622418515355	http://www.ontogen.org/Agriculture/record5081599560713484

Row Count 16

Figure 2.10: Data Properties of 3rd individual under ind1

2.8 Discussion

Currently, open data is gaining momentum as all countries are looking for approaches and methodologies to make their data knolwedgebale to the general public. An ontology framework for New Zealand open data will contribute to the research by opening up paths for data encoding and semantic queries. However, designing such an ontology is quite challenging because of the heterogeneous data sources and multiple formats used to publish them. It is difficult to analyse big data sets and discover relevant data fields to generate the triples. For an initial experiment, data sets from agriculture, land, and rainfall sectors have been encoded and designed in the form of ontologies. This preliminary design would be used as a prototype to encode and link other sectors of the government. Moreover, other countries can take advantage by selecting this design and can semantically link their open government data initiatives. However, it would be

difficult for non-English countries to use this framework if they do not have an English version of the data sources to be linked.

2.9 Conclusion

This paper shows that linked open data has substantially increased in momentum in recent years, with an increasing number of countries opening up their data for public consumption. For instance, the UK government is pioneering the creation of a web of linked open government data, which has opened up the data for the developers so that they can use it for economic and other benefits. Furthermore, in this paper we have also discussed the growth of various OGD initiatives, their benefits and limitations of linked open data. While, open data has several benefits, it also entails numerous barriers in the fields of technology, legislation, use and complexity. Open data has no value by itself, as it only becomes valuable when utilised in an application, primarily for decision support.

There is an abundance of misconceptions about linked open data, such as that all information should be published unrestrictedly and that publishing open data will automatically bring transparency. These misconceptions are used a lot to convince data providers to open up their data to the public. However, it ignores the fact that there are numerous limitations due to the heterogeneous nature of open data. Our analysis of literature showed that OGD supports creative use of government data, which in turn increases governance transparency. Identification of potential OGD datasets also helps public bodies to better understand and manage their own data. It also provides a transparent way of notifying the results of governmental regulations.

Currently, the main focus is on the supply of the data. However, the success of linked open data depends on the quality and use of the data. It is of the utmost importance for governments to work on policies and procedures towards the usage and access of linked

open data. Governments have to take a broader view of having such an infrastructure which would help users access the data and use it for productive purposes. Such policies and infrastructure would promote the current level of engagement of ordinary citizens by increasing transparency and trust. In order to get full benefits from linked open data, it is vital to put data and information together in a context which would create novel knowledge and enable other potential useful applications and services. Linked open data facilitates knowledge and innovation of the nexuses in the data, which is the prime mechanism for information integration and management.

Chapter 3

Ontology-Based Semantic Search Framework for Disparate Datasets

Abstract

The public sector provides open data to create new opportunities, stimulate innovation, and implement new solutions that benefit academia and society. However, open data is usually available in large quantities and often lacks quality, accuracy, and completeness. It may be difficult to find the right data to analyse a target. There are many rich open data repositories, but they are difficult to understand and use because the data can only be used with a complex set of keyword search options, and even then, irrelevant or insufficient data may eventually be retrieved. To alleviate this situation, ontology-based semantic search has been proven to be an effective way to improve the quality of related content queries in such repositories. In this paper, we propose a new method of semantic linking and storing open government datasets of New Zealand's agriculture, land, and rainfall sectors based on the use of ontology. The generated ontology can construct integrated data, in which a unified query can be applied to extract richer and more useful information. To validate our model, we showed how to link ontologies manually and

automatically. Manual linking requires domain experts, and automatic linking reduces the overhead of relying on domain experts to manually link concepts. The results of this method are promising in terms of improving data quality and search efficiency. In the future, the proposed model can be integrated with other domain ontologies.

3.1 Introduction

The main goal of the World Wide Web (WWW) has always been to allow people to easily access information, regardless of whether machines also use the web network to transmit information. The Semantic Web is a web of data that uses Semantic Web standards for annotation (Berners-Lee, Hendler & Lassila, 2001), such as Resource Description Framework (RDF) (Ma, Capretz & Yan, 2016) and Ontology Web Language (OWL) (Hitzler, 2021), which are usually related to each other based on context. We live in a world where data is ubiquitous and integral to our lives and is indispensable as members of organisations and communities (Davenport et al., 2006). The amount of data is growing at an unprecedented rate, and it is believed that the potential growth pattern will continue to rise (Chen, Chiang & Storey, 2012).

Open data can be obtained through a variety of channels, such as open data repositories, portals, websites, and open-source tools. However, the open availability of data does not guarantee the integrity and consistency of the published data. The heterogeneous nature of data makes data extraction lengthy and time-consuming. Alternatively, portals and data on the website are only suitable for keyword-based searches, where the keywords entered by the user match the available data descriptions (Fleiner, 2018). However, the problem is that users are not properly told what to search for and how they can modify keywords to get the best results. In addition, for a better search, the same alternative words may appear, but the user can't know these terms because the user may not be familiar with the structure used by the data publisher to describe it. There

may be synonyms that match the user's intent, but the user does not know the actual terminology the publisher uses to refer to their repository. Semantic search solves this problem and aims to improve the accuracy of the search by considering the searcher's intention and the contextual importance of the search term (Shadbolt, Berners-Lee & Hall, 2006). The motivation of our research is to explore whether semantic technology can improve the usability and efficiency of search in more and more open data.

An effective semantic search engine attempts to analyse the user's intention to search for content and the expected meaning of particular search content. If we link data by analogy, it will help citizens find and use information more easily. The openness of data links allows developers to build useful applications and helps contribute to the development of the country. Entrepreneurs can use connected open data to build innovative business ideas and products that help and stimulate the country's economic growth. Linked open data enhances public participation and indirectly helps governments improve efficiency and decision-making. The government will inform citizens of their behaviour. This will help build trust between the government and the citizens. It will also improve government processes and services. However, making OGD (Ruijer & Meijer, 2020) useful is challenging. Although the number of attempts to disclose OGD is rapidly increasing, it is still a huge challenge to reach the maximum capacity of OGD sources and support the consumption and release of this data by all partners. One of the main challenges in solving this problem is the heterogeneity of the data formats and structures used by government agencies. Due to this heterogeneity, both data providers and data users face technical challenges.

In this article, we introduced a new method for real-world analysis of open data and demonstrated an example of using ontology. Ontologies are semantically related, so SPARQL queries are imposed to extract useful information (Yin, Gromann & Rudolph, 2021). A prototype has been implemented that supports the semantic linking of concepts

related to the agricultural, land and rainfall sector datasets published by various departments of the New Zealand government. Preliminary findings indicate that ontology semantic search can be applied to open data to improve the quality and effectiveness of the search. The SPARQL query applied to the generated ontology will bring information according to the user's request. This means that users no longer need to search for different data sources because all the information is concentrated in one place. We summarise the main contributions of this paper as follows:

- We conducted a study to analyse New Zealand's open data set and convert it into linked open data to extract useful information for novice users and society.
- We developed a prototype to generate the semantical link ontology for the open government datasets automatically. We present a case study-based evaluation where open datasets of agriculture, land, and rainfall sectors are used to generate semantically linked ontologies.
- We designed a SPARQL interface to impose queries on the generated ontologies so that knowledgeable data can be extracted more easily.

The structure of the paper is as follows. Section 3.2 summarises the related work. Section 3.3 provides detailed information on the proposed method of automatically creating and linking semantic ontologies. Section 3.4 describes the results and discussion. In the last section, we concluded.

3.2 Related Work

This section introduces the background of work carried out in the field of open government data implementing ontology frameworks, models, or prototypes. This review aims to compare the proposed solutions and technologies. (Kalampokis et al., 2011) proposed

architecture for determining indirect links between entities to create linked data. The main goal is to create owl: SameAs links between Uniform Resource Identifiers (URIs) to connect different Open Government Data (OGD). In addition, a prototype scene was developed that involved three participants: the school, the Athens Regional Secondary Education Bureau, and the Ministry of Education. To access data consumption related to a particular school, five execution steps were performed. The SPARQL Protocol and RDF Query Language (SPARQL) endpoint are used to access the specified data set. However, the whole process is slow and tedious, requiring users to put in a lot of effort. An open-source framework called Silk was discussed to merge different data sources, but there were no implementation details on how it was implemented. More research is needed to understand the relationship between political priorities and data models. In addition, it is necessary to describe a method or approach to semantically encoding different data sources. The result of the SPARQL query is knowledge data (or linked data), which can provide multiple benefits such as transparency, reusability, economic growth, and public participation (Attard et al., 2015).

In (Fragkou et al., 2016) and (Galiotou & Fragkou, 2013), the author connects the open data project by using an interface based on the E-GIF ontology and the Jena framework. Jena is an open-source semantic web framework. It provides an application programming interface (API) for extracting data from RDF. The SPARQL endpoint is also used to query the data set. However, the focus is on the operation and features of the Jena framework. The proposed method seems to be based on the effectiveness of the tool. In addition, the RDF model is created using a pre-designed E-GIF ontology. There is no clear evidence of the status of the implementation details of the E-GIF ontology. (Jiang et al., 2019) proposed a search engine prototype, which is used to link the concept of transportation domain ontology manually and automatically.

The results proved a higher quality and more effective search for open data. Nevertheless, the proposed system can be enhanced by using additional data sets and provide

more detailed metadata descriptions, which helps to generate semantic link data. At this stage, only one domain is considered, and there is no evidence on how to incorporate other domains into the prototype. In (Escobar et al., 2020), the author proposed a state-of-the-art theory that provides a new method for creating and publishing ontology-based systems by using linked open data from Valencia's water resources management. The proposed ontology can be used to identify the correlation between water sources, leaks, and population. Water resource management decision-making needs to integrate multiple heterogeneous data sources and various data domains. The main goal is to help decision-makers achieve better results by using rich and comprehensive information. However, the interconnection can be enhanced by merging additional data sets from different fields.

With the increase in knowledge representation, deep learning, Natural Language Processing (NLP), machine learning, and daily data volume, ontologies have become more and more important. Ontology engineering is the method and process of research and development of ontology, including the representation, formal naming, and description of categories, properties, and relationships between concepts, data, and entities (Kendall, McGuinness & Ding, 2019). In addition, the implementation of e-government by semantic network technology has brought various types of challenges, including economic, cultural, human, technical, social, data quality, and legislation. By strengthening knowledge sharing, citizens can gain greater advantages from using semantic web applications in e-government (ALSHEHAB[1], ALAZEMI, YOUSEF & ALFAYLY, 2021).

The above-mentioned studies only used open government data from different countries and departments to investigate prototypes, E-GIF ontologies, search engines, and ontology-based frameworks. Although extensive research has been conducted, there is still more room to implement mechanisms and technologies to take advantage of open data procedures to extract valuable information for the benefit of the public. There is an

urgent need for a new method, especially to convert different open data sources into a common form, so that a huge knowledge base can be created, in which multiple data sources can be used to generate semantically rich data. The purpose of this research is to use the open government data set of the New Zealand government to develop a simple and accurate method to generate semantically rich automatic ontologies.

3.3 Methodology

This section describes how to convert comma-separated value (CSV) data to Web Ontology Language (OWL) (Bechhofer et al., 2004). Since it is fully automated and therefore less time-consuming, the proposed method is different from traditional data conversion. It converts the CSV data set to OWL format and allows the generation of semantic links between different ontologies to develop an automated knowledge base that can use SPARQL to query and extract data. The basic concepts of this method are as follows:

- The syntax and semantics of CSV follow the constraints and definitions specified in the RFC4180 document describing the dialect. RFC4180 is used as the dialect description because it can automatically recognise the CSV file format. The default method of the CSV format library is used to parse CSV files.

- The CSV data is annotated using Dublin Core Metadata (Kapidakis, 2020) to achieve interoperability with the OWL metadata vocabulary. It allows accurate and consistent organisation and enrichment of data across multiple modes.

- The union of two ontology files and the HyperSQL(HSQL) (Widianto & Warmayudha, 2020) database is used to semantically link one or more ontology. The union here consists of all ontology classes, individuals, and data properties of

the two ontology files. The main purpose of creating a joint file is to save all the information in the ontology file without losing information.

- HSQL database helps to convert ontology classes into tabular form. This conversion makes the semantic link search operation simple because the tabular form is easy to traverse when performing linear search operations.

- The cosine similarity measure is used to identify the similarity between the literal values of the ontology data properties. It provides a function for comparing two strings and returns the similarity score used to identify the most suitable match for an individual.

- For the ontology merging process, we select common properties from the generated ontology, and generate semantic links based on these properties. This helps to align two or more ontologies into a single modular ontology.

- For ontology visualisation, protection tools are used. In addition, a SPARQL (Pérez, Arenas & Gutierrez, 2009) query is imposed on the ontology, and Apache Jena is used to convert OWL triples into RDF/Turtle form so that SPARQL queries can be executed to obtain the desired results.

3.3.1 Architecture and Process Flow of the Application

A prototype has been developed using the proposed method. We have implemented the prototype of the CSV to OWL conversion and the related conversion mechanism. The conversion method accepts a CSV file and Dublin core meta-words as input and outputs the converted OWL. The created OWL file is located in the local memory of the system. In addition, two or more generated ontologies are semantically linked through the use of union, HSQL, and sequential search operations. To obtain accurate results, the cosine similarity measure is applied to the semantically generated ontology to

eliminate duplication. Then convert the generated OWL file to Turtle format to execute SPARQL queries. In our proposed model, the entire ontology generation process is divided into three stages: 1) use the Protege tool to convert and visualize CSV to OWL; 2) generate semantic links between two or more ontologies; and 3) query the SPARQL of the created ontology interface. In Figure 3.1, we have shown the overall design and process, as well as the key elements of the architecture.

3.3.2 CSV to OWL Conversion and Visualization Using Protege Tool

The first step involves automatically creating an ontology from a CSV data file. Users can choose to upload the CSV file directly or enter the CSV file's accessible Uniform Resource Locator (URL). The current process can only accept CSV data streams as input. However, other modes can be easily added, including portable document format (PDF), hypertext markup language (HTML), keyhole markup language (KML), and JavaScript object notation (JSON).

The Apache Commons CSV library is used to parse CSV files, which follow the constraints and definitions described in the dialect of the RFC4180 document. This specifies the format of the CSV file, such as headers, line endings, and escape characters, and helps with the processing of CSV's text-based fields. The default mode specifies that non-Unicode characters in the CSV file are replaced with Unicode based on the dialect definition. The CSV parser uses various functions to read and parse the rows, cell values, and reference values of CSV data.

The conversion of CSV data to OWL mainly requires the addition of metadata notes describing the data interpretation method. Because it has been recognised as a tool, non-experts can use it to quickly generate transparent and informative records of information resources, while also providing an effective search for resources in an integrated world.

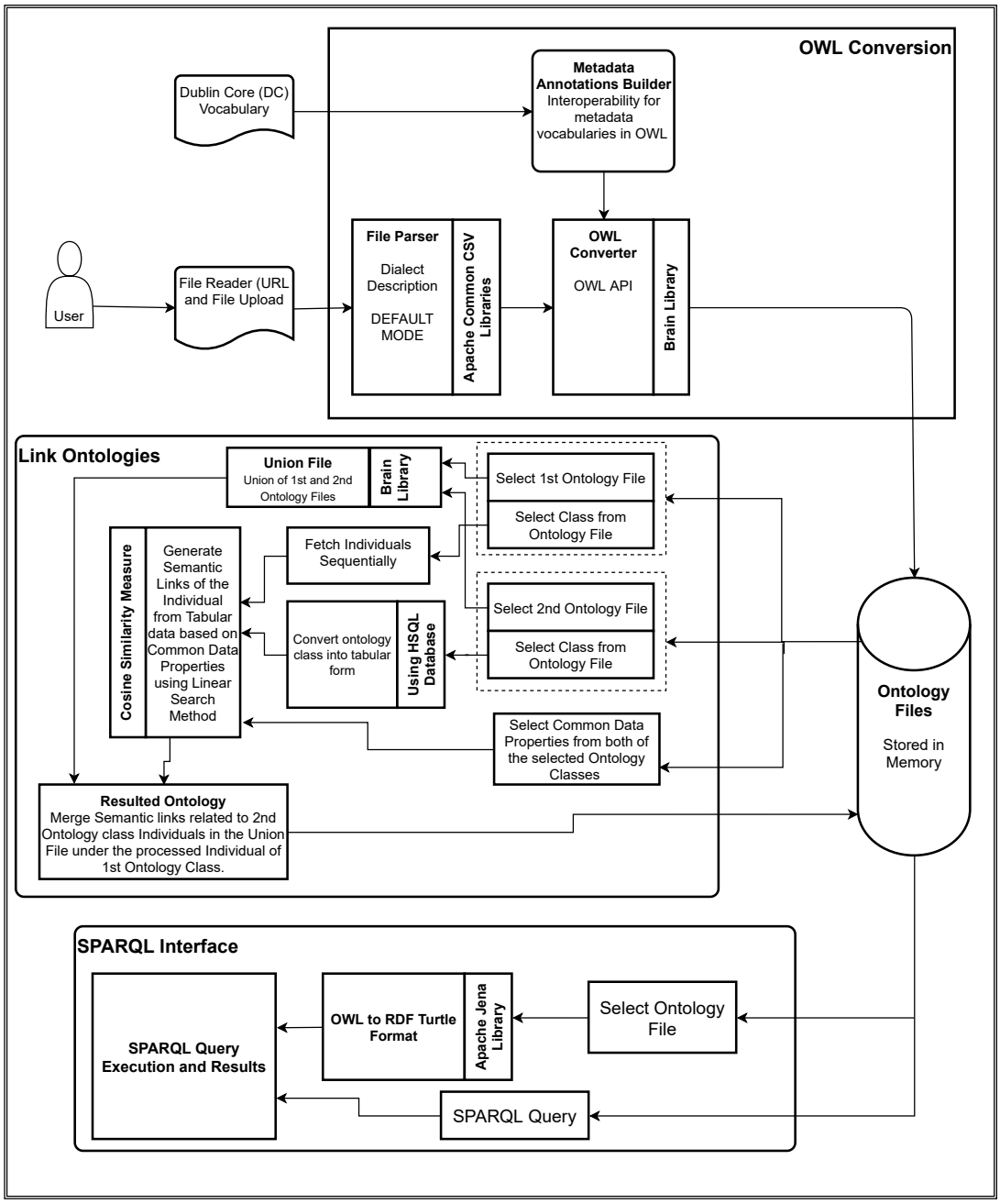


Figure 3.1: Architecture and Process of the Proposed Model

The Dublin Core Resource Description Framework Architecture (RDFS) vocabulary is used to define common metadata. The reason for using the RDFS vocabulary is that it has been widely recognised as a mechanism by which non-experts can easily build accurate and comprehensive records for data sources, and at the same time, they can perform searches on such resources well in an interconnected environment. The name of the CSV file entered by the user is designated as the ontology name, the column header of the CSV file is designated as the data property, and the value stored in the data property is regarded as the entity of the ontology. Each row under the column heading is treated as an individual, and each record is assigned a unique identifier name. While performing all the transformations, this process will also help add corresponding axioms between data properties and individuals to explain the meaning of classes and their relationships. After adding all the axioms, the ontology is stored in local memory, and the generated ontology is visualized using the protege 5.5 tools.

3.3.3 Generating Semantic Links Between Two or More Ontologies

Based on the first step, all ontology files are created, and these files are stored in the system's local storage. To semantically link the ontology, the user must select two or more ontology files. Since the ontology file can have multiple ontology classes, it is necessary to select a specific class from the selected ontology file. The user also needs to select common data properties from the two selected ontology classes to perform semantic link operations. Once the user has generated all the ontologies mentioned in Section 3.2, the user can select two different ontologies to create a semantic link between them. After selecting the two ontology files, the next step is to merge these files to form a union. The union consists of all ontology classes, individuals, and data properties. The main purpose of creating a joint file is to preserve all the information in

the ontology file, so as not to lose any information.

As part of this process, we use the HSQL database to convert the classes of the second ontology file into tabular form. The reason behind this conversion is to make the semantic link search operation simple because the table form is easy to traverse when performing a linear search operation. The next step is to get the individuals one by one sequentially from the first ontology class. Further, the data properties of the first ontology type individual are compared with the tabular data, and a linear search is performed to find the semantically related second ontology type individual. It is challenging to semantically link these data sets based on common properties because individuals may have duplicate values. Therefore, the ideal solution for linking these ontology classes is based on standards that determine the similarity between the literal values of the data properties of the ontology classes. To overcome this challenge, we used the cosine similarity measure, which provides a function for comparing two strings and returning the similarity score.

Since we have repetitions in the area names, cosine similarity is useful for situations where repetition is important. The linear search method will locate individuals in tabular data to identify semantic links on selected public properties. It will continue to browse through each individual of the data in order until it finds a match or searches the entire table. The generated semantic link is further merged into the second ontology individual in the union file. It will lead to a semantically linked ontology of two different data departments or domains. The overall architectural flow of this process is shown in Figure 3.1. In addition, the linking process mentioned here is a way, which means that we can link ontology class 1 to ontology class 2, and vice versa. The two-way link process will be considered in future enhancements.

3.3.4 SPARQL Interface to Query the Generated Ontology

Once we have obtained our semantic link ontology. The third and final stage is to query the generated ontology with the help of the SPARQL interface. The design of the SPARQL interface allows users to select the required ontology file for the query. To query in SPARQL, we need to convert the generated OWL ontology into an RDF format, such as Turtle. Writing SPARQL queries involving complex OWL expressions ranges from challenging to unpleasant because SPARQL query syntax is based on Turtle (Unadkat, 2015), and this does not apply to OWL. SPARQL queries for OWL data must encode the RDF serialisation of OWL expressions: these queries are often lengthy, difficult to write and understand. For the conversion process, we used Apache Jena, which is a Java library that can be used to convert OWL files to RDF Turtle format and provides APIs to query SPARQL from Java applications. Fig. 1 highlights the process of the SPARQL interface.

3.4 Results and Discussion

The above methods have been evaluated using the open datasets of agriculture, land and rainfall on the New Zealand government website. First, we create the ontology of agriculture, land, and rainfall datasets. To do this, we enter the URL of the dataset (https://data.mfe.govt.nz) into the system. The entered URL is parsed using a file parser, and the CSV data set is converted into an OWL file using an OWL converter. These OWL files are stored in the system's local memory. The schematic diagram of the agricultural ontology is shown in Figure 3.2. Agriculture is a subcategory of owl: it has 5 data properties (i.e., area hectares, farm type, FID, region, year) and 620 individuals (from individual 1_159956422651 to individual 602_159956422961). When the mouse pointer hovers over an individual, it will highlight the data property assertion for that

particular individual. Similarly, we create and store ontologies for land and rainfall datasets. In addition, these created ontologies are semantically linked so that SPARQL queries can be applied to extract knowledgeable data. To do this, we select second ontology files from the local drive of the system and manually identify their common data properties.

In the second ontology, year, and region are common data properties. Therefore, these data properties are selected for semantic link generation.We create a joint file by combining two ontology files to avoid any loss of information. The joint file has been put aside and will be merged with the resulting ontology later. For semantic link generation, agriculture and land ontology are regarded as the first and second ontology files, respectively. If the ontology has multiple classes, we need to ensure that the appropriate class is selected from the selected ontology file. From the agriculture OWL file and the land OWL file, we have selected the agriculture and land categories, respectively. The agricultural individuals are obtained one by one, and the land is transformed into a table form, which makes the operation of semantic link search simple.The table form is easy to traverse when performing linear search operations.

Finally, a linear search of agricultural individual and land table data is used to generate semantic links. Here, cosine similarity is used to avoid repetition. Before getting the final linked ontology, we combine the previously created joint file with the generated ontology file, so that we can put semantic and non-semantic data together. Figure 3.3 shows the semantically related ontology of the agricultural and land datasets. The red dotted arrow indicates the semantic connection between land and agricultural individuals. The total number of individuals is 892, and the total number of semantic links is 168. Due to screen size limitations, we cannot display all semantically linked individuals. In order to create the ontology, the existing ontology is not used. All ontologies are newly generated using open datasets from the three departments of the New Zealand government's agriculture, land, and rainfall. The agricultural profile and

all generated ontologies are OWL-RL, because they are mainly for applications that require scalable reasoning without giving up too much representational potential. In addition, it also provides features, specifications, axioms, reasoning, and expressions for the design of RDF triples (W3C, 2017).

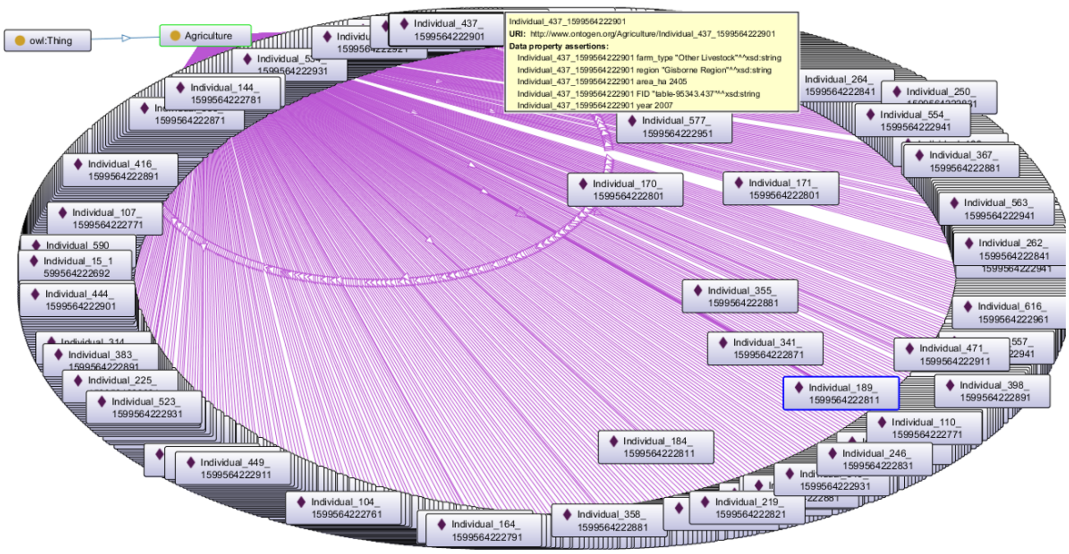


Figure 3.2: Agriculture ontology with all individuals

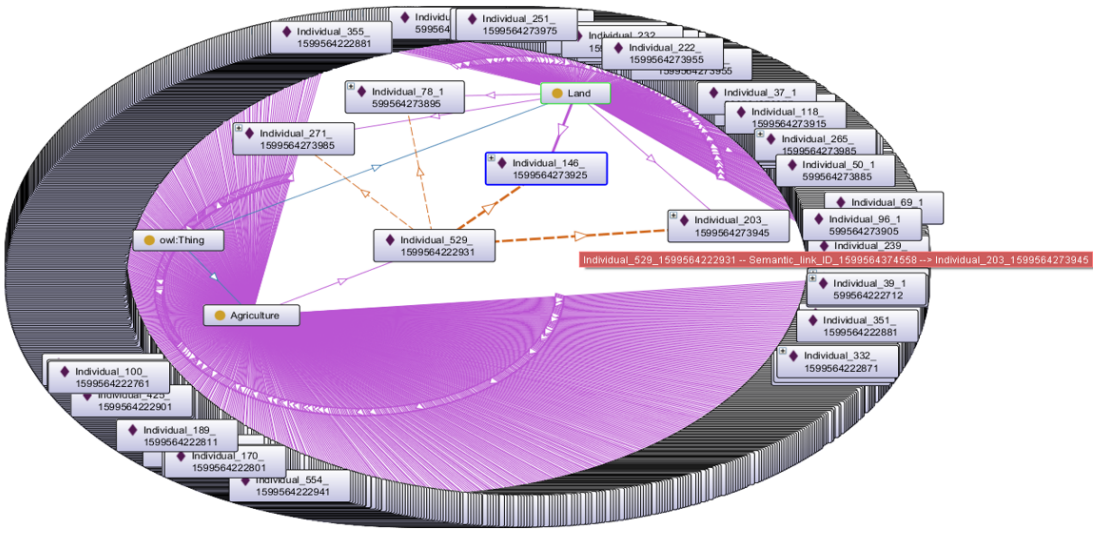


Figure 3.3: Semantically linked ontology of agriculture and land datasets

It is important to analyse the validity of the generated ontology to evaluate our

method. To achieve this, we provide some sample SPARQL queries to the system. We record the responses to these queries and manually evaluate their consistency. SPARQL queries are applied to data sets to retrieve valuable information. Traditional information extraction techniques are very time-consuming because the data sets must be thoroughly read, but the RDF query language ultimately makes the task simple because all data is stored as triples. By using SPARQL queries, you can easily find knowledge-rich data. For testing purposes, we implemented the following SPARQL queries on the generated ontology:

- Query 1: Find the area hector used by dairy and exotic forest in the year 2012 in the Auckland region.
- Query 2: Find the area hector used by exotic grasslands in Gisborne in the year 2008 and what is the rainfall rate for that year and region.

We show the syntax of SPARQL queries 1 and 2 in Figures 3.4 and 3.5, respectively.

```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT (SUM(?area) AS ?total_area)
WHERE
{{
?ind1 rdf:type owl:NamedIndividual .
?ind1 rdf:type <http://www.ontogen.org/Land/Land>.
?ind1 <http://www.ontogen.org/Land/area_ha>?area .
?ind1 <http://www.ontogen.org/Land/year>2012 .
?ind1 <http://www.ontogen.org/Land/region>"Auckland" .
?ind1 <http://www.ontogen.org/Land/type>"exotic_forest" }
UNION {
?ind1 rdf:type owl:NamedIndividual .
?ind1 rdf:type <http://www.ontogen.org/Agriculture/Agriculture>.
?ind1 <http://www.ontogen.org/Agriculture/area_ha>?area .
?ind1 <http://www.ontogen.org/Agriculture/year>2012 .
?ind1 <http://www.ontogen.org/Agriculture/region>"Auckland Region" .
?ind1 <http://www.ontogen.org/Agriculture/farm_type>"Dairy".
}}
```

Figure 3.4: Structure of SPARQL Test Query 1

Compared with the traditional relational database query, the SPARQL query is simple

```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT (SUM(?area) AS ?total_area) ?Rainfall ?Year ?site
WHERE
{
?ind1 rdf:type owl:NamedIndividual .
?ind1 rdf:type <http://www.ontogen.org/Land/Land>.
?ind1 <http://www.ontogen.org/Land/area_ha>?area .
?ind1 <http://www.ontogen.org/Land/year>2008 .
?ind1 <http://www.ontogen.org/Land/region>"Gisborne" .
?ind1 <http://www.ontogen.org/Land/type>"exotic_grassland" .
?ind2 rdf:type <http://www.ontogen.org/Rainfall/Rainfall>.
?ind2 <http://www.ontogen.org/Rainfall/Year>2008 .
?ind2 <http://www.ontogen.org/Rainfall/site>"Gisborne" .
?ind2 <http://www.ontogen.org/Rainfall/r95ptot>?Rainfall .
?ind2 <http://www.ontogen.org/Rainfall/Year>?Year .
?ind2 <http://www.ontogen.org/Rainfall/site>?site .
}
```

Figure 3.5: Structure of SPARQL Test Query 2

but efficient as shown in Figures 3.4 and 3.5. SPARQL can be used to query any database, and any middleware can be used to interpret the results as RDF. On the other hand, relational database requests are limited to specific databases. SPARQL is a Hypertext Transfer Protocol (HTTP) protocol that allows connection to any SPARQL endpoint through the structured transport layer. We process the test queries (1 and 2) and capture their results, as shown in Figures 3.6 and 3.7, respectively. For consistency and accuracy, we conduct manual analysis. Therein, it is found that the result captured by the system is accurate.

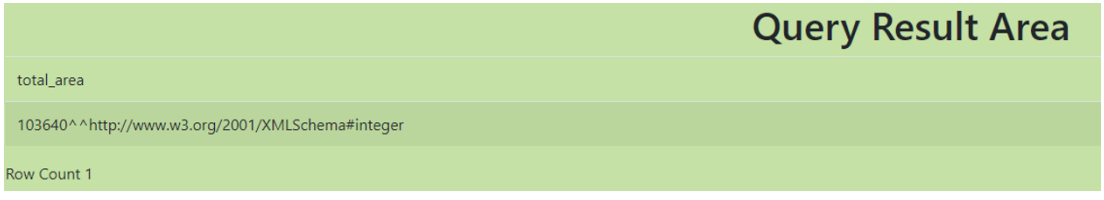


Figure 3.6: Result of the Test Query 1 for Agriculture and Land Dataset Ontology

At present, the New Zealand open data set does not have a semantically related ontology framework. It is necessary to access information through an open portal.

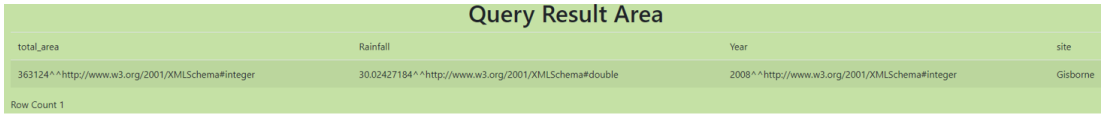


Figure 3.7: Result of the Test Query 2 for Land and Rainfall Dataset Ontology

Due to the heterogeneity of the open data set, it is difficult for end users to find useful information without traversing multiple pages or links. Even in most cases, they ultimately have no information. The proposed system helps to improve data quality and search efficiency in less time, because users can easily find linked information in one place, and by applying a query, the required data can be extracted.

It is expected that the introduction of automatic ontology generation prototypes will improve the consumption and access of open data sources by the New Zealand government. This method currently only considers the CSV data set and needs to be further improved, such as by trying to use other formats and using two-way semantic encoding. The results are positive, and it is worthwhile to further study the use of this method to connect to other areas of government. This method has practical uses in applications such as generating ontologies for any open data set of the New Zealand government. However, the ontology generation of semantic links requires some manual analysis, and users need to identify common properties between different data sets so that semantic links can be generated.

In addition, the protege tool is used to generate the visualisation of the ontology. It has plug-ins that can accommodate the visualisation of ontology's classes, instances, and data properties. Users can visualise part of the entire ontology. However, when the scale of data grows, the protege cannot handle it, and it becomes challenging to visualize the generated ontology. The large size of the ontology also affects the overall performance and loading time of the knowledge base. However, visualization is not the focus of research, but in the future, visualization can be improved by using third-party plug-ins or creating an additional tool that can improve the overall visualization of the

ontology. The implemented SPARQL interface processes the imposed queries, which are limited to users with technical knowledge. At the current stage of research, natural language queries that can help non-technical and non-SPARQL users are not considered. These functions can be implemented in future enhancements of the proposed system. In addition, the current model is validated using SPARQL endpoints, where example queries are used to generate knowledge-rich data. No other reasoners or editors are used at this stage.

A key advantage of this research is that it will automate the ontology generation process, and semiautomatic methods will help generate semantic links. These findings indicate that automatic ontology generation and semantic encoding of different data sets may be useful tools for leveraging valuable knowledge. Despite the success, an important limitation is that semantic links can only be generated in one way. The links work from left to right. For example, if we want to semantically link agricultural and land datasets, we need to keep the agricultural ontology on the left. In this way, we will get the ontology of agriculture -> land semantic association. If we want to go the other way, we need to keep the land body on the left. In the future, the limitations of this research must be considered to generate two-way semantic links in one go. In addition, the semantic link generation method can be fully automated in the future. To this end, a process can be implemented that can parse the data set and record public entries, and then the table can be used to generate semantic links.

3.5 Conclusion

In this paper, we propose a new method of using ontology semantics to link and store open government datasets of New Zealand's agriculture, land, and rainfall sectors. Our comprehensive approach includes the following: 1) The process of reading and parsing the CSV dataset; 2) The conversion process of converting the CSV dataset

in OWL using Dublin Core metadata annotations; 3) the realisation of the process of generating semantic links between two or more generated ontologies; 4) A SPARQL interface development is used to query and extract useful information from the generated semantically rich ontologies. To validate the proposed solution, the OWL conversion process is used to convert various existing CSV files collected from the New Zealand govt.nz data portal. Our solution can generate an automatic ontology of any data set owned by the New Zealand government. However, the semantic link generation process requires some manual work in identifying the common properties between ontology files. In the future, we plan to consider the conversion of other formats and implementing an algorithm that can simplify the process of semantic link generation. In addition, our case study implementation proves the effectiveness and feasibility of our proposed method. Automatic ontology generation and semantic coding of different data sets can be useful tools for utilising valuable knowledge. Future work includes improving the visualisation and processing time of the generated ontology. In addition, the proposed framework can be used to test open data sets from other countries and fields. In order to allow non-technical users to access the system, natural language query capabilities can be added to future enhancements to the system.

Chapter 5

Conclusion

5.1 Introduction

This chapter presents the conclusion of the thesis along with the limitations and future directions of the proposed framework. Section 5.2 provides a summary of the thesis contribution. Furthermore, section 5.3 presents the limitations of this thesis, followed by future research recommendations.

5.2 Findings

The most significant observation of this study is that an ontology framework is implemented using a multi-layer approach to accommodate the framework's reliability, usability, and efficiency. A layered architecture enables the entire system to grow and expand as independent modules. The primary research question is, "Can disparate structured data published by various government departments be computationally with RDF encoding and semantically linked in an ontology framework?" The first experiment creates an ontology to transform open data into linked open data. New Zealand's open data sets of agriculture, land, and rainfall sectors are encoded as ontologies. This

prototype can encode and connect other government sectors. Other countries can benefit by semantically linking their open government data. However, non-English countries can't adapt to this model. Our most intriguing finding is that the implementation of prototypes for automatic ontology generation will lead to an increase in the number of open data sources utilised by government agencies. This strategy only considers the CSV data set; therefore, it needs to be improved by exploring the possibility of utilising other formats and using two-way semantic encoding. The findings are significant, and it would be beneficial to conduct additional research on utilising this technology to link data of other government departments and agencies. Ontologies can be generated for various domains using this method for any publicly available data set that is managed by any government or nation worldwide.

The second research question is, "Can semantically linked data with RDF encoding be made available using SPARQL endpoints to satisfy the requirements of a wide range of stakeholders and the general public?" Accessing useful data is crucial because it will help the stakeholders and business owners to make essential decisions. As a part of the layered approach, a SPARQL endpoint is implemented, and a set of queries are imposed on the generated semantic ontologies to extract valuable information. Despite the success demonstrated, a significant limitation is that the ontology development of semantic linkages necessitates some manual analysis, and users must find common properties among distinct data sets for semantic link generation. Future research should therefore seek to address this issue by imposing algorithms or metrics to automate the process. Moreover, one of the time-consuming and laborious processes is the visualisation of the generated ontology. We utilised a third-party plugin called protégé to visualise the generated ontology. However, the available plugins of protégé are not reliable and unable to deal with a large ontology. Ontology visualisation is computationally expensive and time-consuming hence, we suggest that future studies should examine the visualisation of ontologies.

5.3 Summary

Open data presents both benefits and problems for research and innovation in the areas of knowledge and system design. The opportunities for application developers and researchers to acquire and integrate publicly available data from a variety of sources are remarkable. Many open data platforms, however, are designed in such a manner that it is challenging, at some point impossible, to derive value from data. Open data, especially government data, continues to get a lot of attention because it has the potential to give tech-savvy citizens more power, change how governments should work, and make public services better.

Data is one of the vital aspects of our day-to-day lives. These days, everything on the web is governed by data. However, presently, data is facing the challenges of being machine unreadable and inaccessible. Several efforts have been made to view and resolve the structured format issues of data, but most of the time, the semantic relationships between the data are often not considered. The ability to reuse and enrich data through the use of external sources is made possible by linked data, which also makes it possible to extend data models. As a direct result of this, putting data together and figuring out how it all fits together becomes both more effective and easier. In a general sense, various LOD repositories have lately come into existence, which is significant in terms of data access and integration. However, there is still a limited capacity regarding particular sectors when it comes to reusing the datasets. This is due to a variety of factors, such as the use of standard formats (such as PDF) that are not easily accessible in any way. In this regard, the usage of LOD that is gradually rising may be of assistance in the process of integrating various datasets. It's also important to note that using public SPARQL endpoints allows the company's internal information to be combined with information from external repositories, but it requires IT skills to do so.

The information available in the open government initiatives is ineffective for sharing and reuse due to the enormous volume and heterogeneity. The World Wide Web has connected the world through hyperlinks between web documents. These hyperlinks are used to navigate free text between HTML pages. Data can thus be accessed via a single web link that integrates other connections into the original page (Bizer et al., 2011). The increased growth of a knowledgeable social system has resulted in people becoming more aware of their rights and wanting to be involved and know about governance strategies. This has led to the need for transparency by governing bodies, resulting in the release of data to the public to make the transparency rationale for policy decisions.

This thesis has proposed a framework to transform open data into linked open data to generate ontology by using the open government datasets of the New Zealand government. The data sets of three sectors (agriculture, land, and rainfall) of the NZ government are used to analyse the ontology generation process. A SPARQL interface is used to extract knowledgeable data from the ontology. Furthermore, an expert evaluation was conducted, which evaluated the efficiency and effectiveness of the proposed framework.

The ontology framework uses open government data extracted from the New Zealand government portal as input. It is organised using four layers: the data conversion layer, the ontology generation layer, the semantic link generation layer, and the SPARQL interface layer. The overall architecture is designed in a layered fashion for several reasons, such as to facilitate flexibility, maintainability, scalability, and to accommodate efficiency, reliability, and usability. Unlike traditional approaches, a layered architecture enables the entire system to grow and expand as independent modules. The full design expands and remains synchronised because the frameworks will adjust without restrictions based on each module's specifications. The primary reason behind the architectural design was to enable the reuse of the various components of the framework

by multiple applications. For example, some external systems may be interested in using only the ontology generation module, which is easily accomplished due to the architecture's stand-alone nature.

Initially, the data conversion module will run and convert the input datasets into the desired OWL (Ontology Web Language) form so that OWL triples can be created, visualised, and queried as ontologies. The data conversion module is independent, whereas the ontology generation module is dependent on it. Further, the semantically linked ontology module is dependent on the output of the ontology generation module. The last module, i.e., the SPARQL interface, is dependent on the output ontology, which can either be a single ontology, or a semantically linked ontology.

Any of these modules can only be used again after the data conversion module has been completed and accessed. We have explained each module and its connectivity with its corresponding modules to demonstrate the overarching operation of the ontology framework. The data conversion layer is responsible for managing the information sources that are used by the framework, which are essentially open government data sources. Data conversion is a crucial component of the system and is accessed in smaller components of varying complexity. The modules are explored in depth in chapter 3 to show how each part of the system accesses data. At this stage, only CSV datasets are used. Other formats such as PDF, XML, JSON, HTML, and KML, etc., are not considered. This is due to two reasons: Firstly, CSV files are plain text files, making them easier to import into a spreadsheet or other storage database. Secondly, it is useful for better organising large volumes of data. Moreover, the CSV syntax and semantics adhere to the constraints and definitions of the RFC4180 document's dialect description. The CSV Format Library's DEFAULT method is used to parse the CSV files.

The CSV data is accompanied by annotations of Dublin Core metadata, which provides interoperability for the OWL metadata vocabulary. The Dublin Core metadata allows the accuracy, organisation, and enrichment of knowledge in various schemes.

Additionally, the framework makes use of a variety of built-in libraries to facilitate parsing and conversion processes. The data conversion layer comprises the input to the framework, which either uses the CSV data files, or the URL of the data files based on the users' preference. The data files are extracted from the open data government portal of New Zealand. The ontology generation layer receives input from the data conversion layer in the form of the parsed CSV file and processes it to generate OWL triples, which are stored in the local memory of the system, and visualised as an ontology by using the protégé tool. Protégé is an ontology editing tool, developed by Standford University. It is free software with both a graphical user interface (GUI) and an application programming interface (API), which gives it a great deal of adaptability. One of the tools that is used the most frequently for ontology editing is protégé.

Data sources from the New Zealand government and Dublin Core vocabulary are used as input sources for the framework. The reason behind using Dublin core vocabulary was that the core schema of Dublin is a compact collection of vocabulary definitions that can be used to characterise a variety of different types of resources. Dublin Core Metadata can be used for a variety of things, including basic resource descriptions, merging metadata vocabularies from various metadata standards, and ensuring metadata vocabularies in Linked Data cloud and semantic web implementations are interoperable.

In this thesis, the results of all ontology generation using different data sets have been thoroughly presented and discussed. Various queries are considered to analyse the correlation between the datasets used to generate domain ontologies. The primary investigations performed for open data transformation and ontology generation give substantial evidence for the proposed framework's effectiveness, usefulness, and efficiency. These contributions are discussed in chapters 2-4 of this thesis.

An initial part of this thesis investigated open data initiatives and the role of open data in transparency and how open data availability can impact the overall growth of a

country. In chapter 2, we analysed the literature around linked open data tools, models, architectures, algorithms, and frameworks. We have also analysed and discussed the benefits and limitations of open data. The analysis identified research gaps, which prompted the development of an effective and efficient ontology framework to transform open data into linked open data.

In chapter 3, the architecture of the proposed framework is presented. The overall architecture is built in layers, to accommodate efficiency, reliability, and usability. Unlike traditional approaches, a layered architecture enables the entire system to grow and expand as independent modules. The full design expands and remains synchronised because the frameworks will adjust without restrictions based on each module's specifications. The primary reason behind the architectural design was to enable the reuse of the various components of the framework by multiple applications. For example, some external systems may be interested in using only the ontology generation module, which is easily accomplished due to the architecture's stand-alone nature.

In the first layer, data is collated from data portals; as an experiment, we have utilised the open data sets of three sectors: agriculture, land, and rainfall of the New Zealand (NZ) government. All selected datasets are initially converted into CSV format to generate triples to design the ontology. The generated CSV datasets for agriculture, land, and rainfall sectors are parsed, and OWL triples are developed to create each sector's ontologies. As a result, individual ontologies of agriculture, land, and rainfall are generated.

In the second layer, the individual ontologies of agriculture, land, and rainfall domains are used to generate the semantically linked ontologies. To generate the semantic links among the ontologies, common properties are identified. Whereas all other individuals are also collated as union files, none of the triples is left behind. To avoid duplicating the individuals' values, cosine similarity is used at this stage. Once the semantically linked ontologies are generated, they will be stored on the local drives

of the system.

In the third and final layer, disparate data sources are published via the SPARQL interface to extract valuable information by imposing appropriate SPARQL queries. To query in SPARQL, it is required to convert the generated OWL ontologies into a resource description framework (RDF) format such as Turtle. The whole conversion process is supported by Apache Jena, which has provided the required application programming interfaces (APIs) to query SPARQL. Furthermore, in this chapter, SPARQL queries are imposed on the framework to extract useful information and to check the consistency of the system. The CSV to OWL conversion used here is different as compared to the normal data conversion as it considers the RDF triples and we are converting the CSV datasets into OWL triples so that we can form an ontology. OWL vocabularies are the most updated ones. It also helps to generate semantically enriched data when multiple ontologies are linked. To query an ontology, it should be in turtle form. OWL ontologies do not support the turtle format. To impose the queries, the OWL ontologies are converted to RDF/turtle format.

In chapter 4, the evaluation of the ontology framework to transform open data into linked open data is presented. The evaluation of the framework is divided into two aspects: written response and scale-based response. Firstly, we have conducted an expert evaluation where experts selected from the industry have profound experience, and a written set of questions with predefined criteria is provided to the experts to answer. Secondly, the experts have given a five-point scale response based on the described criteria of usability, reliability, correctness, and effectiveness (efficiency) of the proposed ontology framework. The feedback is gathered and analysed in two ways: the scale-based response is captured and analysed in the form of charts where the rating for usability, reliability, correctness, effectiveness, and efficiency of the framework is identified and analysed based on the response of the experts. For the written feedback, NVivo is used to study the responses of the experts so that relevant and useful points can

be captured. The results identified by using NVivo show that most of the experts agree with the completeness, effectiveness, and appropriateness of the proposed framework. However, few improvements and recommendations have been provided by the experts to enhance the framework for the real world. We have incorporated the feasible suggestions of the experts and put some aside for future improvements.

5.4 Limitations and Future Research Recommendations

This section highlights the future research directions and limitations of the proposed framework. The limitations and future directions indicated in this section do not affect the validity of the research contributions provided in this thesis. These, however, urge future studies to advance and enhance the proposed framework and provide a wide range of open data transformation applications.

An ontology framework to transform open data into linked open data will contribute to research by allowing data encoding and semantic queries. However, creating such an ontology is rather difficult due to the diverse data sources and numerous formats utilised to release them. It is challenging to analyse large data sets and find important data fields in order to generate triples. In the first experiment, data sets from the agriculture, land, and rainfall sectors were encoded and constructed in the form of ontologies. This preliminary concept would be used as a prototype to encode and connect other government sectors. Furthermore, other countries can benefit from this approach by semantically linking their open government data. Non-English countries, on the other hand, would find it impossible to adopt this paradigm.

The implementation of automatic ontology generation prototypes is expected to increase government consumption and access to open data sources. This approach presently only considers the CSV data set and should be enhanced by attempting to use other formats and employing two-way semantic encoding. The results are promising,

and it is useful to conduct additional research on the usage of this technology to link to other departments of government. This approach can be used to generate ontologies for any open data set maintained by any government or country. However, the ontology development of semantic linkages necessitates some manual analysis, and users must find common properties among distinct data sets for semantic link generation.

In addition, the protege tool is utilised to generate the ontology visualisation. It includes plug-ins that allow you to see the classes, instances, and data attributes of an ontology. The user can see a portion of the whole ontology. When the amount of data is too large for the protégé to handle, it becomes difficult to visualise the resulting ontology. The ontology's large size has an impact on the knowledge base's overall performance and processing time. However, while visualisation is not the focus of this study, it can be improved in the future by leveraging third-party plug-ins or developing a new tool to improve the overall visualisation of the ontology.

The outlined SPARQL interface executes the enforced queries, which are only accessible to technical users. Natural language queries that can assist non-technical and non-SPARQL users are not explored at this stage in this thesis. These features could be included in future versions of the proposed framework. Furthermore, the current model is evaluated using SPARQL endpoints, which create knowledge-rich data using sample queries. At this point, no further reasoners or editors are employed. In the future, our objective is to provide parallel processing in terms of the computation of the results of a single query across partitions where the query match exists.

One of the most significant benefits of this research is that it will automate the ontology generation process, with semiautomatic approaches assisting in the generation of semantic linkages. These findings suggest that automatic ontology development and semantic encoding of various data sets could be beneficial tools for utilising important information. Despite the success, there is one significant drawback: semantic links can only be formed in one way. The links work from left ontology to right. For instance,

if we would like to semantically link two ontologies such as agriculture and land, we need to retain the agriculture ontology on the left. As a result, you will have the semantically linked ontology of agriculture -> land. If we wish to go the other way around, we need to keep land ontology on the left. In the future, the limits of this research must be acknowledged in order to generate two-way semantic linkages in a single step. Furthermore, the semantic link generation process can be totally automated in the future. For this purpose, a process can be implemented that can parse the datasets and record the common entries, and later that table can be used to generate the semantic links.

This thesis demonstrates our approach that automatically develops an ontology for any government data source. The semantic link generation procedure, on the other hand, necessitates some manual work in detecting common features between ontology files. In the future, we plan to investigate converting other formats and developing an algorithm to speed up the process of generating semantic links. We also plan to enhance the current methodologies in order to develop a new schema and classification of semantic collections of data from various ontologies. To do this, we will make use of more expressive ontology techniques such as ontology matching and ontology learning.

References

Agrawal, S., Deshmukh, J., Srinivasa, S., Jog, C., Bhavaani, K. & Dhek, R. (2013). A survey of indian open data. In *Proceedings of the 5th ibm collaborative academia research exchange workshop* (pp. 1–4). doi: https://doi.org/10.1145/2528228.2528230

Alani, H., Brewster, C. & Shadbolt, N. (2006). Ranking ontologies with aktiverank. In *International semantic web conference* (pp. 1–15).

Alexopoulos, C., Spiliotopoulou, L. & Charalabidis, Y. (2013). Open data movement in greece: a case study on open government data sources. In *Proceedings of the 17th panhellenic conference on informatics* (pp. 279–286). doi: https://doi.org/10.1145/2491845.2491876

Al-Khalifa, H. S. (2013). A lightweight approach to semantify saudi open government data. In *2013 16th international conference on network-based information systems* (pp. 594–596). doi: https://doi.org/10.1109/NBiS.2013.99

ALSHEHAB[1], A., ALAZEMI, N., YOUSEF, M. & ALFAYLY, A. (2021). Challenges of applying semantic web approaches on e-government web services: Survey. *International Journal*, *10*(2). doi: 10.30534/ijatcse/2021/1041022021

Altayar, M. S. (2018). Motivations for open data adoption: An institutional theory perspective. *Government Information Quarterly*, *35*(4), 633–643. doi: https://doi.org/10.1016/J.GIQ.2018.09.006

Apuke, O. D. (2017). Quantitative research methods: A synopsis approach. *Kuwait Chapter of Arabian Journal of Business and Management Review*, *33*(5471), 1–8. doi: 10.12816/0040336

Arcelus, J. (2012). Framework for useful transparency websites for citizens. In *Proceedings of the 6th international conference on theory and practice of electronic governance* (pp. 83–86). doi: https://doi.org/10.1145/2463728.2463749

Attard, J., Orlandi, F. & Auer, S. (2016). Value creation on open government data. In *2016 49th hawaii international conference on system sciences (hicss)* (pp. 2605–2614). doi: https://doi.org/10.1109/HICSS.2016.326

Attard, J., Orlandi, F., Scerri, S. & Auer, S. (2015). A systematic review of open government data initiatives. *Government information quarterly*, *32*(4), 399–418. doi: https://doi.org/10.1016/j.giq.2015.07.006

Azevedo, P. C. N., Pinto, V. A., Bastos, G. S. & Parreiras, F. S. (2015). Using linked open data in geographical information systems. In *International conference on geographical information systems theory, applications and management* (pp.

152–166). doi: 10.1007/978-3-319-29589-3_10

Bahanshal, A. O. & Al-Khalifa, H. S. (2013). Toward recipes for arabic dbpedia. In *Proceedings of international conference on information integration and web-based applications & services* (pp. 331–335). doi: https://doi.org/10.1145/2539150.2539199

Baiyang, L. & Ruhua, H. (2016). A study on the approaches of value realization of open government data.

Barati, M., Bai, Q. & Liu, Q. (2017). Mining semantic association rules from rdf data. *Knowledge-Based Systems*, *133*, 183–196. doi: https://doi.org/10.1016/j.knosys.2017.07.009

Bauer, F. & Kaltenböck, M. (2011). Linked open data: The essentials. *Edition mono/monochrom, Vienna*, *710*.

Bechhofer, S., Van Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D. L., Patel-Schneider, P. F., ... others (2004). Owl web ontology language reference. *W3C recommendation*, *10*(2), 1–53.

Berners-Lee, T. (2009). *Linked data.* Retrieved from `https://www.w3.org/DesignIssues/LinkedData.html`

Berners-Lee, T., Hendler, J. & Lassila, O. (2001). The semantic web. *Scientific american*, *284*(5), 34–43.

Bertot, J. C., Jaeger, P. T. & Grimes, J. M. (2010). Using icts to create a culture of transparency: E-government and social media as openness and anti-corruption tools for societies. *Government information quarterly*, *27*(3), 264–271. doi: https://doi.org/10.1016/j.giq.2010.03.001

Bilgin, G., Dikmen, I. & Birgonul, M. T. (2014). Ontology evaluation: An example of delay analysis. *Procedia Engineering*, *85*, 61–68. doi: https://doi.org/10.1016/j.proeng.2014.10.529

Bizer, C., Heath, T. & Berners-Lee, T. (2011). Linked data: The story so far. In *Semantic services, interoperability and web applications: emerging concepts* (pp. 205–227). IGI global. doi: 10.4018/978-1-60960-593-3.ch008

Bojārs, U. & Liepiņš, R. (2014). The state of open data in latvia: 2014. *arXiv preprint arXiv:1406.5052*. doi: https://doi.org/10.48550/arXiv.1406.5052

Boyatzis, R. E. (1998). *Transforming qualitative information: Thematic analysis and code development*. sage.

Brank, J., Grobelnik, M. & Mladenic, D. (2005). A survey of ontology evaluation techniques. In *Proceedings of the conference on data mining and data warehouses (sikdd 2005)* (pp. 166–170).

Brank, J., Mladenic, D. & Grobelnik, M. (2006). Golden standard based ontology evaluation using instance assignment. In *Eon@ www.*

Breitman, K., Salas, P., Casanova, M. A., Saraiva, D., Gama, V., Viterbo, J., ... Chaves, M. (2012). Open government data in brazil. *IEEE Intelligent Systems*, *27*(03), 45–49. doi: 10.1109/MIS.2012.25

Brewster, C., Alani, H., Dasmahapatra, S. & Wilks, Y. (2004). Data driven ontology evaluation.

Burton-Jones, A., Storey, V. C., Sugumaran, V. & Ahluwalia, P. (2005). A semiotic metrics suite for assessing the quality of ontologies. *Data & Knowledge Engineering*, *55*(1), 84–102.

Casellas, N. (2009). Ontology evaluation through usability measures. In *Otm confederated international conferences" on the move to meaningful internet systems"* (pp. 594–603). doi: 10.1007/978-3-642-05290-3_73

Charalabidis, Y., Alexopoulos, C. & Loukis, E. (2016). A taxonomy of open government data research areas and topics. *Journal of Organizational Computing and Electronic Commerce*, *26*(1-2), 41–63. doi: https://doi.org/10.1080/10919392.2015.1124720

Chen, H., Chiang, R. H. & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS quarterly*, *36*(4), 1165–1188. doi: https://doi.org/10.2307/41703503

Consoli, S., Presutti, V., Recupero, D. R., Nuzzolese, A. G., Peroni, S., Gangemi, A. et al. (2017). Producing linked data for smart cities: The case of catania. *Big Data Research*, *7*, 1–15. doi: https://doi.org/10.1016/j.bdr.2016.10.001

Corrêa, A. S., Corrêa, P. L. P. & da Silva, F. S. C. (2014). Transparency portals versus open government data: An assessment of openness in brazilian municipalities. In *Proceedings of the 15th annual international conference on digital government research* (pp. 178–185). doi: https://doi.org/10.1145/2612733.2612760

Coyle, K. (2012). Semantic web and linked data. *Library technology reports*, *48*(4), 10–14.

Cypress, B. (2018). Qualitative research methods: A phenomenological focus. *Dimensions of Critical Care Nursing*, *37*(6), 302–309. doi: 10.1097/DCC.0000000000000322

Dahbi, K. Y., Lamharhar, H. & Chiadmi, D. (2018). Toward an evaluation model for open government data portals. In *International conference europe middle east & north africa information systems and technologies to support learning* (pp. 502–511).

data.govt.nz. (2020). *Open government information and data programme.* Retrieved from `https://www.data.govt.nz/standards-and-guidance/open-data/open-data-nz/`

Davenport, T. H. et al. (2006). Competing on analytics. *Harvard business review*, *84*(1), 98.

Dawes, S. S. & Helbig, N. (2010). Information strategies for open government: Challenges and prospects for deriving public value from government transparency. In *International conference on electronic government* (pp. 50–60). doi: 10.1007/978-3-642-14799-9_5

D.Beckett, T.-L. (2020). *Turtle-terse rdf triple language-w3c team submission.* Retrieved from `https://www.w3.org/TeamSubmission/2011/SUBM-turtle-20110328/`

DiFranzo, D., Graves, A., Erickson, J. S., Ding, L., Michaelis, J., Lebo, T., ... others (2011). The web is my back-end: Creating mashups with linked open government data. In *Linking government data* (pp. 205–219). Springer. doi: https://doi.org/

10.1007/978-1-4614-1767-5_10

Elliott-Mainwaring, H. (2021). Exploring using nvivo software to facilitate inductive coding for thematic narrative synthesis. *British Journal of Midwifery*, *29*(11), 628–632.

Escobar, P., Roldán-García, M. d. M., Peral, J., Candela, G. & Garcia-Nieto, J. (2020). An ontology-based framework for publishing and exploiting linked open data: A use case on water resources management. *Applied Sciences*, *10*(3), 779. doi: 10.3390/app10030779

Fernández-López, M. & Gómez-Pérez, A. (2002). The integration of ontoclean in webode..

Fleiner, R. (2018). Linking of open government data. In *2018 ieee 12th international symposium on applied computational intelligence and informatics (saci)* (pp. 1–5). doi: https://doi.org/10.1109/SACI.2018.8441014

Foundation, W. W. W. (2018). *Open data barometer.* Retrieved from `https://opendatabarometer.org/doc/leadersEdition/ODB-leadersEdition-Report.pdf`

Fragkou, P., Galiotou, E. & Matsakas, M. (2014). Enriching the e-gif ontology for an improved application of linking data technologies to greek open government data. *Procedia-Social and Behavioral Sciences*, *147*, 167–174. doi: 10.1016/j.sbspro.2014.07.141

Fragkou, P., Kritikos, N. & Galiotou, E. (2016). Querying greek governmental site using sparql. In *Proceedings of the 20th pan-hellenic conference on informatics* (pp. 1–6). doi: https://doi.org/10.1145/3003733.3003807

Galiotou, E. & Fragkou, P. (2013). Applying linked data technologies to greek open government data: a case study. *Procedia-social and behavioral sciences*, *73*, 479–486. doi: https://doi.org/10.1016/j.sbspro.2013.02.080

Gangemi, A., Catenacci, C., Ciaramita, M. & Lehmann, J. (2006). Modelling ontology evaluation and validation. In *European semantic web conference* (pp. 140–154). doi: https://doi.org/10.1007/11762256_13

Gardner, S. P. (2005). Ontologies and semantic data integration. *Drug discovery today*, *10*(14), 1001–1007.

González, J. C., Garcia, J., Cortés, F. & Carpy, D. (2014). Government 2.0: a conceptual framework and a case study using mexican data for assessing the evolution towards open governments. In *Proceedings of the 15th annual international conference on digital government research* (pp. 124–136). doi: https://doi.org/10.1145/2612733.2612742

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge acquisition*, *5*(2), 199–220.

Gruzd, A. & Roy, J. (2014). Investigating political polarization on twitter: A canadian perspective. *Policy & internet*, *6*(1), 28–45. doi: https://doi.org/10.1002/1944-2866.POI354

Guarino, N., Oberle, D. & Staab, S. (2009). What is an ontology? In *Handbook on ontologies* (pp. 1–17). Springer. doi: 10.1007/978-3-540-92673-3_0

Harrison, T. M., Guerrero, S., Burke, G. B., Cook, M., Cresswell, A., Helbig, N., ...

Pardo, T. (2012). Open government and e-government: Democratic challenges from a public value perspective. *Information polity*, *17*(2), 83–97.

Harrison, T. M. & Sayogo, D. S. (2014). Transparency, participation, and accountability practices in open government: A comparative study. *Government information quarterly*, *31*(4), 513–525. doi: https://doi.org/10.1016/J.GIQ.2014.08.002

Hartmann, J., Spyns, P., Giboin, A., Maynard, D., Cuel, R., Suárez-Figueroa, M. C. & Sure, Y. (2005). D1. 2.3 methods for ontology evaluation. *EU-IST Network of Excellence (NoE) IST-2004-507482 KWEB Deliverable D*, *1*.

Hashemi, P., Khadivar, A. & Shamizanjani, M. (2018). Developing a domain ontology for knowledge management technologies. *Online Information Review*. doi: https://doi.org/10.1108/OIR-07-2016-0177

Haslhofer, B. & Isaac, A. (2011). data. europeana. eu: The europeana linked open data pilot. In *International conference on dublin core and metadata applications* (pp. 94–104).

Hassanzadeh, O. (2011). Introduction to semantic web technologies & linked data. *University of Toronto*.

Hassanzadeh, O. & Consens, M. P. (2009). Linked movie data base. In *Ldow*.

Heath, T. & Bizer, C. (2011). Linked data: Evolving the web into a global data space. *Synthesis lectures on the semantic web: theory and technology*, *1*(1), 1–136. doi: https://doi.org/10.2200/S00334ED1V01Y201102WBE001

Hendler, J., Holm, J., Musialek, C. & Thomas, G. (2012). Us government linked open data: semantic. data. gov. *IEEE Intelligent Systems*, *27*(03), 25–31. doi: 10.1109/MIS.2012.27

Hitzler, P. (2021). A review of the semantic web field. *Communications of the ACM*, *64*(2), 76–83. doi: https://doi.org/10.1145/3397512

Hitzler, P., Krotzsch, M. & Rudolph, S. (2009). *Foundations of semantic web technologies*. Chapman and Hall/CRC. doi: https://doi.org/10.1201/9781420090512

Hlomani, H. & Stacey, D. (2014). Approaches, methods, metrics, measures, and subjectivity in ontology evaluation: A survey. *Semantic Web Journal*, *1*(5), 1–11.

Hoxha, J., Brahaj, A. & Vrandečić, D. (2011). Open. data. al: increasing the utilization of government data in albania. In *Proceedings of the 7th international conference on semantic systems* (pp. 237–240). doi: https://doi.org/10.1145/2063518.2063558

Index, G. O. D. (2014). *Place overview*. Retrieved from `https://index.okfn.org/place/`

International, O. K. (April,2017). *The comprehensive knowledge archive network*. Retrieved from `https://ckan.org/`

Janssen, M., Charalabidis, Y. & Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. *Information systems management*, *29*(4), 258–268. doi: https://doi.org/10.1080/10580530.2012.716740

Jiang, S., Hagelien, T. F., Natvig, M. & Li, J. (2019). Ontology-based semantic search for open government data. In *2019 ieee 13th international conference on semantic computing (icsc)* (pp. 7–15). doi: 10.1109/ICOSC.2019.8665522

Jørgensen, K. B. & Jensen, L. E. (2011). *Introduction to nvivo 9.0.* Aarhus: Aarhus University.

Kalampokis, E., Tambouris, E. & Tarabanis, K. (2011). A classification scheme for open government data: towards linking decentralised data. *International Journal of Web Engineering and Technology*, *6*(3), 266–285. doi: https://doi.org/10.1504/IJWET.2011.040725

Kalampokis, E., Tambouris, E. & Tarabanis, K. (2013). Linked open government data analytics. In *International conference on electronic government* (pp. 99–110). doi: 10.1007/978-3-642-40358-3_9

Kapidakis, S. (2020). Consistency and interoperability on dublin core element values in collections harvested using the open archive initiative protocol for metadata harvesting. In *Keod* (pp. 181–188). doi: 10.5220/0010112001810188

Kaur, P. & Nand, P. (2021a). Implementing automatic ontology generation for the new zealand open government data: An evaluative approach. In *International conference on advances in computing and data sciences* (pp. 26–36). doi: https://doi-org.ezproxy.aut.ac.nz/10.1007/978-3-030-81462-5_3

Kaur, P. & Nand, P. (2021b). Towards transparent governance by unifying open data. *IAENG International Journal of Computer Science*, *48*(4).

Kaur, P., Nand, P., Naseer, S., Gardezi, A. A., Alassery, F., Hamam, H., ... Shafiq, M. (2022). Ontology-based semantic search framework for disparate datasets. *INTELLIGENT AUTOMATION AND SOFT COMPUTING*, *32*(3), 1717–1728. doi: 10.32604/iasc.2022.023063

Kendall, E., McGuinness, D. & Ding, Y. (2019). Ontology engineering (synthesis lectures on the semantic web: Theory and technology). *Morgan & Claypool, San Rafael*, 1–136. doi: https://doi.org/10.2200/S00834ED1V01Y201802WBE018

Khusro, S., Jabeen, F., Mashwani, S. R. & Alam, I. (2014). Linked open data: towards the realization of semantic web-a review. *Indian Journal of Science and Technology*, *7*(6), 745.

Kim, J. & Storey, V. C. (2011). Construction of domain ontologies: Sourcing the world wide web. *International Journal of Intelligent Information Technologies (IJIIT)*, *7*(2), 1–24.

Klein, R. H., Klein, D. B. & Luciano, E. M. (2018). Open government data: concepts, approaches and dimensions over time. *Revista economia & gestão*, *18*(49), 4–24. doi: https://doi.org/10.5752/P.1984-6606.2018V18N49P4-24

Klyne, G. (2004). Resource description framework (rdf): Concepts and abstract syntax. *http://www. w3. org/TR/2004/REC-rdf-concepts-20040210/*.

Kontokostas, D., Bratsas, C., Auer, S., Hellmann, S., Antoniou, I. & Metakides, G. (2011). Towards linked data internationalization-realizing the greek dbpedia.

Koznov, D., Andreeva, O., Nikula, U., Maglyas, A., Muromtsev, D. & Radchenko, I. (2016). A survey of open government data in russian federation. In *Proceedings of the international joint conference on knowledge discovery, knowledge engineering and knowledge management* (p. 173–180). Setubal, PRT: SCITEPRESS - Science and Technology Publications, Lda. Retrieved from `https://doi.org/10.5220/0006049201730180` doi: 10.5220/0006049201730180

Kucera, J. & Chlapek, D. (2014). Benefits and risks of open government data. *Journal of Systems Integration*, *5*(1), 30. doi: 10.20470/jsi.v5i1.185

Kucera, J., Chlapek, D., Klímek, J. & Necaskỳ, M. (2015). Methodologies and best practices for open data publication. In *Dateso* (pp. 52–64).

Kučera, J., Chlapek, D. & Nečaskỳ, M. (2013). Open government data catalogs: Current approaches and quality perspective. In *International conference on electronic government and the information systems perspective* (pp. 152–166).

Le, N.-T., Ichise, R. & Le, H.-B. (2010). Detecting hidden relations in geographic data. In *Proceedings of the 4th international conference on advances in semantic processing* (pp. 61–68). doi: 10.1.1.460.3211

Liu, Q., Bai, Q., Ding, L., Pho, H., Chen, Y., Kloppers, C., ... others (2011). Linking australian government data for sustainability science-a case study. In *Linking government data* (pp. 181–204). Springer. doi: 10.1007/978-1-4614-1767-5_9

Ltd, Q. I. P. (2020). *Nvivo.* Retrieved from `https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/home`

Ma, Z., Capretz, M. A. & Yan, L. (2016). Storing massive resource description framework (rdf) data: a survey. *The Knowledge Engineering Review*, *31*(4), 391–413. doi: https://doi.org/10.1017/S0269888916000217

Máchová, R. & Lněnička, M. (2017). Evaluating the quality of open data portals on the national level. *Journal of theoretical and applied electronic commerce research*, *12*(1), 21–41. doi: https://doi.org/10.4067/S0718-18762017000100003

Matheus, R., Ribeiro, M. M. & Vaz, J. C. (2012). New perspectives for electronic government in brazil: the adoption of open government data in national and subnational governments of brazil. In *Proceedings of the 6th international conference on theory and practice of electronic governance* (pp. 22–29). doi: https://doi.org/10.1145/2463728.2463734

Mekhabunchakij, K. (2016). Towards modeling linked open data for decision support: An example application of thailand tourism linked data visualization. In *2016 management and innovation technology international conference (miticon)* (pp. MIT–88). doi: https://doi.org/10.1109/MITICON.2016.8025240

Mockus, M. & Palmirani, M. (2017). Legal ontology for open government data mashups. In *2017 conference for e-democracy and open government (cedem)* (pp. 113–124). doi: 10.1109/CeDEM.2017.25

Mutuku, L. N. & Colaco, J. (2012). Increasing kenyan open data consumption: A design thinking approach. In *Proceedings of the 6th international conference on theory and practice of electronic governance* (pp. 18–21). doi: https://doi.org/10.1145/2463728.2463733

Neuendorf, K. A. (2018). Content analysis and thematic analysis. In *Advanced research methods for applied psychology* (pp. 211–223). Routledge. doi: https://doi.org/10.4324/9781315517971

Ngomo, A.-C. N., Auer, S., Lehmann, J. & Zaveri, A. (2014). Introduction to linked data and its lifecycle on the web. In *Reasoning web international summer school* (pp. 1–99). doi: 10.1007/978-3-642-39784-4_1

Nikiforova, A. & McBride, K. (2021). Open government data portal usability: A

user-centred usability analysis of 41 open government data portals. *Telematics and Informatics*, *58*, 101539. doi: https://doi.org/10.1016/J.TELE.2020.101539

Nowell, L. S., Norris, J. M., White, D. E. & Moules, N. J. (2017). Thematic analysis: Striving to meet the trustworthiness criteria. *International journal of qualitative methods*, *16*(1), 1609406917733847. doi: https://doi.org/10.1177/1609406917733847

NZ, G. (2015). *Open government data programme.* Retrieved from `https://www.data.govt.nz/standards-and-guidance/open-data/open-data-nz/`

Obrst, L., Ashpole, B., Ceusters, W., Mani, I., Ray, S. & Smith, B. (2007). *The evaluation of ontologies: Toward improved semantic interoperability.[versión electrónica].* Semantic Web Revolutionizing Knowledge in the Life Science. Springer US.

Obrst, L., Gruninger, M., Baclawski, K., Bennett, M., Brickley, D., Berg-Cross, G., ... others (2014). Semantic web and big data meets applied ontology. *Applied Ontology*, *9*(2), 155–170.

Orszag, P. R. (2009). *Open government directive.* Retrieved from `https://obamawhitehouse.archives.gov/open/documents/open-government-directive`

Paintal, S. (2004). Doing qualitative research in education settings. *Childhood Education*, *80*(3), 166–167.

Pak, J. & Zhou, L. (2009). A framework for ontology evaluation. In *Workshop on e-business* (pp. 10–18). doi: 10.1007/978-3-642-17449-0_2

Paliouras, G., Spyropoulos, C. D. & Tsatsaronis, G. (2011). *Knowledge-driven multimedia information extraction and ontology evolution: bridging the semantic gap* (Vol. 6050). Springer.

Palmirani, M., Martoni, M. & Girardi, D. (2014). Open government data beyond transparency. In *International conference on electronic government and the information systems perspective* (pp. 275–291). doi: 10.1007/978-3-319-10178-1_22

Parundekar, R., Knoblock, C. A. & Ambite, J. L. (2010). Linking and building ontologies of linked data. In *International semantic web conference* (pp. 598–614). doi: https://doi.org/10.1007/978-3-642-17746-0_38

Patel, A. & Jain, S. (2021). Present and future of semantic web technologies: a research statement. *International Journal of Computers and Applications*, *43*(5), 413–422. doi: https://doi.org/10.1080/1206212X.2019.1570666

Patton, M. Q. (2008). *Utilization-focused evaluation*. Sage publications.

Pérez, J., Arenas, M. & Gutierrez, C. (2009). Semantics and complexity of sparql. *ACM Transactions on Database Systems (TODS)*, *34*(3), 1–45. doi: https://doi.org/10.1145/1567274.1567278

Petrov, O., Gurin, J. & Manley, L. (2016). Open data for sustainable development. doi: https://doi.org/10.1596/24017

Quarati, A. & De Martino, M. (2019). Open government data usage: a brief overview. In *Proceedings of the 23rd international database applications & engineering*

symposium (pp. 1–8). doi: https://doi.org/10.1145/3331076.3331115

Refaeilzadeh, P., Tang, L., Liu, H., Liu, L. & Özsu, M. (2009). Encyclopedia of database systems. In *Cross-validation* (pp. 532–538). Springer.

Ronzhin, S., Folmer, E. & Lemmens, R. (2018). Technological aspects of (linked) open data. In *Open data exposed* (pp. 173–193). Springer.

Ruijer, E. & Meijer, A. (2020). Open government data as an innovation process: Lessons from a living lab experiment. *Public Performance & Management Review*, *43*(3), 613–635. doi: https://doi.org/10.1080/15309576.2019.1568884

Sabou, M. & Fernandez, M. (2012). Ontology (network) evaluation. In *Ontology engineering in a networked world* (pp. 193–212). Springer. doi: 10.1007/978-3-642-24794-1_9

Saxena, S. & Muhammad, I. (2018). The impact of open government data on accountability and transparency. *Journal of Economic and Administrative Sciences*. doi: 10.1108/JEAS-05-2017-0044

Shadbolt, N., Berners-Lee, T. & Hall, W. (2006). The semantic web revisited. *IEEE intelligent systems*, *21*(3), 96–101. doi: 10.1109/MIS.2006.62

Sotiriadou, P., Brouwers, J. & Le, T.-A. (2014). Choosing a qualitative data analysis tool: A comparison of nvivo and leximancer. *Annals of Leisure Research*, *17*(2), 218–234. doi: 10.1080/11745398.2014.902292

Sowe, S. K. & Zettsu, K. (2015). Towards an open data development model for linking heterogeneous data sources. In *2015 seventh international conference on knowledge and systems engineering (kse)* (pp. 344–347). doi: 10.1109/KSE.2015.56

Storey, V. C. & Thalheim, B. (2017). Conceptual modeling: enhancement through semiotics. In *International conference on conceptual modeling* (pp. 182–190).

Subedi, R., Nyamasvisva, T. E. & Pokharel, M. (2021). The movement of open government data: A systematic review. *Korea*, *71*(95), 64.

Supekar, K. (2005). A peer-review approach for ontology evaluation. In *8th int. protege conf* (pp. 77–79).

Tan, H., Adlemo, A., Tarasov, V. & Johansson, M. E. (2017). Evaluation of an application ontology. In *Proceedings of the joint ontology workshops 2017 episode 3: The tyrolean autumn of ontology bozen-bolzano, italy, september 21–23, 2017* (Vol. 2050).

Tartir, S., Arpinar, I. B., Moore, M., Sheth, A. P. & Aleman-Meza, B. (2005). Ontoqa: Metric-based ontology quality analysis.

Tartir, S., Arpinar, I. B. & Sheth, A. P. (2010). Ontological evaluation and validation. In *Theory and applications of ontology: Computer applications* (pp. 115–130). Springer. doi: 10.1007/978-90-481-8847-5_5

Taychatanompong, A. & Vatanawood, W. (2019). Sales forecasting using ontology. In *Proceedings of the international multiconference of engineers and computer scientists, pp416-420.*

Terry, G., Hayfield, N., Clarke, V. & Braun, V. (2017). Thematic analysis. *The SAGE handbook of qualitative research in psychology*, *2*, 17–37. doi: https://dx.doi.org/10.4135/9781526405555.n2

Theocharis, S. & Tsihrintzis, G. A. (2014). Ontology development to support the open public data-the greek case. In *Iisa 2014, the 5th international conference on information, intelligence, systems and applications* (pp. 385–390). doi: 10.1109/IISA.2014.6878820

Theocharis, S. A. & Tsihrintzis, G. A. (2013). Open data for e-government the greek case. In *Iisa 2013* (pp. 1–6). doi: 10.1109/IISA.2013.6623722

Ubaldi, B. (2013). Open government data: Towards empirical analysis of open government data initiatives.
doi: https://doi.org/10.1787/19934351

UK, G. (2013). *Open government partnership.* Retrieved from `https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/207772/Open_Data_Charter.pdf`

Unadkat, R. (2015). Survey paper on semantic web. *International Journal of Advanced Pervasive and Ubiquitous Computing (IJAPUC)*, *7*(4), 13–17. doi: 10.4018/978-1-5225-5191-1.ch007

Vafopoulos, M. N., Meimaris, M., Papantoniou, A., Anagnostopoulos, I., Alexiou, G., Avraam, I., ... Loumos, V. (2012). Public spending: Interconnecting and visualizing greek public expenditure following linked open data directives. *Available at SSRN 2064517*. doi: http://dx.doi.org/10.2139/ssrn.2064517

Vicknair, C., Macias, M., Zhao, Z., Nan, X., Chen, Y. & Wilkins, D. (2010). A comparison of a graph database and a relational database: a data provenance perspective. In *Proceedings of the 48th annual southeast regional conference* (pp. 1–6). doi: https://doi.org/10.1145/1900008.1900067

Vrandečić, D. (2009). Ontology evaluation. In *Handbook on ontologies* (pp. 293–313). Springer. doi: 10.1007/978-3-540-92673-3_13

W3C. (2017). *What is rdf triplestore?* Retrieved from `https://ontotext.com/knowledgehub/fundamentals/what-is-rdf-triplestore/`

Walsham, G. (2001). Knowledge management:: The benefits and limitations of computer systems. *European management journal*, *19*(6), 599–608. doi: 10.1016/S0263-2373(01)00085-8

Washington, D. O. (2011). *Open data charter.* Retrieved from `https://www.opengovpartnership.org/`

Widianto, S. R. & Warmayudha, I. P. E. (2020). Hsql database. *Jurnal Mantik*, *4*(3), 1717–1721.

Wong, L. (2008). Data analysis in qualitative research: A brief guide to using nvivo. *Malaysian family physician: the official journal of the Academy of Family Physicians of Malaysia*, *3*(1), 14.

Yang, T.-M., Lo, J., Wang, H.-J. & Shiang, J. (2013). Open data development and value-added government information: Case studies of taiwan e-government. In *Proceedings of the 7th international conference on theory and practice of electronic governance* (pp. 238–241). doi: https://doi.org/10.1145/2591888.2591932

Yang, T.-M. & Wu, Y.-J. (2016). Examining the socio-technical determinants influencing government agencies' open data publication: A study in taiwan. *Government*

Information Quarterly, *33*(3), 378–392. doi: 10.1016/j.giq.2016.05.003

Yin, X., Gromann, D. & Rudolph, S. (2021). Neural machine translating from natural language to sparql. *Future Generation Computer Systems*, *117*, 510–519. doi: https://doi.org/10.48550/arXiv.1906.09302

Zhang, H., Li, Y.-F. & Tan, H. B. K. (2010). Measuring design complexity of semantic web ontologies. *Journal of Systems and Software*, *83*(5), 803–814.

Zhao, L. & Ichise, R. (2014). Ontology integration for linked data. *Journal on Data Semantics*, *3*(4), 237–254. doi: https://doi.org/10.1007/s13740-014-0041-9

Zinke, F. (2009). *Relational database: A practical foundation.* Citeseer.

Zong, N., Nam, S., Eom, J.-H., Ahn, J., Joe, H. & Kim, H.-G. (2015). Aligning ontologies with subsumption and equivalence relations in linked data. *Knowledge-Based Systems*, *76*, 30–41. doi: 10.1016/j.knosys.2014.11.022

Zouaq, A. & Nkambou, R. (2009). Evaluating the generation of domain ontologies in the knowledge puzzle project. *IEEE Transactions on knowledge and data engineering*, *21*(11), 1559–1572. doi: 10.1109/TKDE.2009.25

Zuiderwijk, A. & Janssen, M. (2014). Open data policies, their implementation and impact: A framework for comparison. *Government information quarterly*, *31*(1), 17–29. doi: https://doi.org/10.1016/j.giq.2013.04.003

Zuiderwijk, A. & Janssen, M. (2015). Participation and data quality in open data use: Open data infrastructures evaluated. In *Proceedings of the 15th european conference on e-government* (pp. 351–359).

Zuiderwijk, A., Janssen, M., Choenni, S., Meijer, R. & Alibaks, R. S. (2012). Socio-technical impediments of open data. *Electronic Journal of e-Government*, *10*(2), pp156–172.

Appendix A

Exceptions to Activities requiring AUTEC approval (6)

The following activities do not require AUTEC approval:

6.7. Where a professional or expert opinion is sought, except where this is part of a study of the profession or area of expertise.

More Details can be found at: https://www.aut.ac.nz/research/researchethics/guidelines-and-procedures#6