Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

School of Computer Science and Statistics

The Ethical Dangers of Public Data

Harry Thompson

May 2, 2022

A dissertation submitted in partial fulfilment
of the requirements for the degree of
MAI (Computer Engineering)

Declaration

I hereby declare that this dissertation is entirely my own work and that it has not been submitted as an exercise for a degree at this or any other university.

I have read and I understand the plagiarism provisions in the General Regulations of the University Calendar for the current year, found at `http://www.tcd.ie/calendar`.

I have also completed the Online Tutorial on avoiding plagiarism 'Ready Steady Write', located at `http://tcd-ie.libguides.com/plagiarism/ready-steady-write`.

Signed: *Harry Thompson* Date: *02/05/22*

Abstract

The Ethical Dangers of Public Data

Harry Thompson

University of Dublin, Trinity College, 2022

Supervisor: Prof. Declan O'Sullivan

Governments around the world are increasingly making available datasets on a variety of topics affecting their citizens. Throughout the Covid-19 pandemic governments have released case numbers and vaccination records. This paper explores Linking public datasets together, and investigating how Linked Open Data could potentially affect people's lives. An open data linking application was developed to form a concrete use case to answer & explore the research question: *"To what extent can public datasets be linked together in what could be considered an ethically questionable matter ?"*. To investigate the Potential ethical issues that could arise with public data. Three datasets were linked in the areas of: Housing prices, Social demographics, & Vaccination statistics. These datasets were linked together using freely available software to: Convert CSV data to an RDF triplestore, using an R2RML mapping based on a bespoke ontology. The RDF triples from the mapping were utilized to create a knowledge graph in GraphDB, and a set of competence questions in SPARQL were used to explore the connections between the datasets with a User Interface based in python. The links between the datasets were used to explore data ethics & public data, their uses and potential effects on wider society. It is clear that as society progresses to be increasingly data driven, we need to be aware of data ethics, and its role in our society.

Acknowledgements

Thanks so much to Declan O'Sullivan, my project supervisor, without his help this project would not have been possible .

Furthermore, I would like to express my sincere gratitude to my parents and friends for their unwavering support and encouragement as I continued my research and writing this dissertation.

Contents

List of Figures

List of Tables

Nomenclature

CSV	Comma Separated Values
GDRP	General Data Protection Regulation
HTTP	Hypertext transfer protocol
LOD	Linked Open Data
OWL	Web Ontology Language
RDF	Resource description framework
R2RML	Rdb to Rdf Mapping Language
SPARQL	SPARQL Protocol and RDF Query Language
URI	Uniform resource identifiers
W3C	World Wide Web Consortium
XML	Extensible Markup Language

1 Introduction

1.1 Motivation

Public datasets are being released more and more by governments around the world in today's modern world. The Open data strategy by the Irish government (3) is an example of one such initiative. The National data Strategy (4) in the UK is another example of a governmental open data initiative.

The idea behind Open Data is to make data stored by government agencies easily accessible online for reuse and redistribution. Fundamentally with any published open datasets, the basic right to the protection of confidential personal and commercial data will not be harmed by the publication of open data, and any material that could allow an individual entity (person or business) to be identified shall be excluded in these situations. However, this project will investigate if this is in fact the case, and the potential risks of public data.

Nonetheless, it is important to note the potential positive impact of public data, for example a political benefit of the release of increased amounts of governmental Open Data allows for greater transparency and trust in government, as well as the possibility of increased efficiencies through better data governance. In terms of economic benfits, according to some studies, Open Data has a significant economic potential. According to McKinsey(5), a worldwide market powered by Open Data from all sectors would generate an additional $3 trillion to $5 trillion of potential revenue every year. In general business, Open Data could have a positive impact in three major areas: business innovation, company creation, and business efficiency Researchers and enterprises will be able to build on government research with easier and faster access to data. This might help increase innovation in industries like drugs and renewable energy. Opportunities for Open Data-inspired businesses or services to add value to data generated by government agencies may arise. Open Data could help businesses and government agencies acquire more precise and full insights into customer preferences and wants, allowing them to be more efficient in meeting those demands while also contributing to smart growth. The construction of Smart Cities to something as simple as better journey times and passenger convenience via the use of real-time passenger information for public transportation or mapping applications are all examples of social and personal benefits.

Despite these many potential positive effects of Open Data, it is important to be aware of the potential negative effects as it is not yet clear if open data initiatives are truly delivering on their promises. According to Gurstein (6) a negative impact of open data is that it empowers only those who are already 'empowered.'This refers to those who have the ability to make efficient use of open data because they have access to open data infrastructures, hardware, software, financial and educational resources, and the skills to do so. Other research has found that measuring the impact and value of open data is extremely difficult, and that impact and value can only be quantified over time. Because open data is such a new phenomena, little is known about its impact and usefulness. Despite the fact that open data may have the potential to have great value and influence, this has yet to be fully verified, and this is mostly speculation.

According to Zuiderwijk (7) potential risks of open data are: Risk of violating legislation by opening data, meaning many datasets cannot be released for legal reasons for example datasets that contain privacy-identifying variables, (policy) sensitive variables, and datasets that have been developed by several organizations with varying levels of security, rules, and legal requirements. Publication of this type of data would result in unfavorable outcomes, as it would be against the law (for example, the data protection law or GDPR) and could hurt the reputation of the entity that provides the data. Another potential open data difficulty involves Difficulties with data ownership: Many government agencies also maintain data from other organizations. Because the governmental agency does not actually own this data, it cannot be made public. Therefore, data that has been created with the help of other organizations cannot be made public. A major area where open data can be negative is where privacy can be violated unintentionally: much effort is taken to remove privacy sensitive variables from datasets so that they can be accessible. As a result, one would believe that privacy is assured. However, by merging data from other sources, it may still be able to track an individual's identity, particularly when open data is paired with social media data. Some further potential negatives of Open data are: Published data can be biased, Misinterpretation and misuse, inadequate consideration for public value and solving societal problems, and Unclear responsibility and accountability. (7)

1.2 Research Question

According to the motivations the research question is addressed as: "What are the potential ethical dangers of Public data, and how can they be addressed ?"

1.2.1 Definition of Terms:

"Linked data"

Linked Data is a set of design principles for sharing machine-readable interlinked data on the Web.

"Open Data"

Open data is data that anyone can access, use and share. It can be exploited, editable or shared by anyone for any purpose, even commercially.

"Big Data"

Big data refers to data sets that are too large or complex to be dealt with by traditional data-processing application software

"Linked Open Data"

Linked Open Data is a way of publishing structured data that allows metadata to be connected and enriched, so that different representations of the same content can be found, and links made between related resources. Linked Open Data is a powerful combination of Linked Data and Open Data in that it is both linked and based on open source.

"Ontology"

An Ontology encompasses a representation, formal naming, and definition of the categories, properties, and relations between the concepts, data, and entities that substantiate one, many, or all domains of discourse

"Data Ethics"

Data ethics is a branch of ethics that evaluates data practices—collecting, generating, analyzing and disseminating data.

1.3 Research Objectives & Goals

Following research, objectives are derived from the research question:

RO1: Survey of state of the art on Open data, data ethics and linked data. The objective is to review existing literature and publicly available tools to determine up to date research and work in the field of Open data, ethics and any other work which has been done considering ethical concerns in linking datasets and the approaches used by them.

RO2: Obtaining datasets which could potentially be linked together in what could be considered an ethically questionable manner This objective requires an approach of gathering datasets which contain similarities which can be connected and used to link them. This is to form a use case to investigate the ethics of linked data.

RO3: Developing an approach to link the datasets together. This objective requires investigating some method to link some common factors in the datasets together, using publicly available tools.

RO4: Developing a tool which would display the some "Competency questions" online to query the linked database. This objective is to demonstrate how the potentially ethically questionable linked datasets could be made publicly available.

RO5: Evaluate the results and potential ethical and societal impacts.

1.4 Contribution

The primary expected contributions of this work are the development of a concrete use case to illustrate the dangers of linking public datasets.

This contribution is intended to benefit the general public by considering the potential dangers of public datasets, in particular those who could be negatively affected by the misuse of such data.

1.5 Structure of the rest of the dissertation

Related background and State-of-the-Art: This chapter discusses Open Data, Data Ethics, semantic web, and general background in the literature.

Design: This chapter discusses the identified requirements for linking the datasets and building the user interface, the functional architecture, design challenges encountered and how they have been addressed.

Implementation: This chapter includes the programming infrastructure used to build the Linked open data application, the implementation and architecture of the different phases of the application, the ontology and the competency questions which have been used to derive the SPARQL queries.

Result and Evaluation: This chapter includes the experiment motivation, goal and hypothesis, experiment setup, discussion about the input datasets, competency questions, experiment results, and observations for the experiment and the overall research.

Conclusion: This chapter includes a discussion about meeting the research objectives, the contributions, future work and final remarks regarding the research.

2 Related Background & State of the Art

2.1 Related Background:

This section discusses about the ethics and data, GDPR and semantic web and Application of ontologies.

2.1.1 Semantic Web:

To understand the concept of linked data, it is important to appreciate how the ideas of linking data began in the literature. One of the first papers to establish this concept was "The semantic web" by Tim Berners-Lee (8). This paper was foundational in establishing the concept of how knowledge is expressed on the web to enchance the capabilities of what was able to be achieved using the data from various websites. Berners-Lee established the concepts of Expressing Meaning (bringing structure to the meaningful content of Web pages), Knowledge Representation (adding structured collections of information and sets of inference rules for computers to use), & Ontologies (the name for a taxonomy and a set of inference rules between datasets) amongst many other important concepts. The Semantic Web is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation. The goal of the Semantic Web is basically to enable machines to comprehend semantic documents and data, not human speech and writings. In essence, The Semantic Web is a technology for sharing data, just as the hypertext Web is for sharing documents. These concepts were revolutionary at the time (2001), and have been built upon by much of the other literature in the area.

2.1.2 Linked Data:

The concept of linked data is closely related to the Semantic Web, and linked data are considered a brain child of the Sematic Web. Bizer, Berners-Lee et al. defined Linked Data as a set of best practices for publishing and connecting structured data on the Web.(9). Haslhofer and Isaac described that linked data is a data publishing technique used on the World Wide Web to connect related data for the purpose of accessibility on the Web. The concept of linked data is closely related to the Semantic Web, and linked data is considered an evolution of the Sematic Web. (10). Linked data is based on hypertext transfer protocol (HTTP), uniform resource identifiers (URI), and Semantic Web standards like the resource description framework (RDF), which allows related data to be linked to one other and to other related resources.

Berners-Lee (11) described four rules for the creation of linked data on the Web:

1. Use URIs as names for things
2. Use HTTP URIs so that people can look up those names
3. When someone looks up a URI, provide useful information using

   the standards (RDF, SPARQL)
4. Include links to other URIs so that they can discover more things

These rules are expectations of behavior for how to how to publish data.

The first rule is to identify things or data objects with URIs. The Uniform Resource Identifier (URI) is a single global identification system for granting unique names to anything, from Web-based digital material to physical objects and abstract concepts. We can use URIs to distinguish between various items or to know if something in one dataset is the same as something in another.

The second rule serves as to give a standard form to allow ease of use, and ensure alternate URIs such as XML tags aren't used. As the HTTP protocol provides a simple way for getting resources, anything that can be recognized by URIs and used in combination with it become easy to find. This facilitates the publication of any type of data and its inclusion in the global data space. (12)

The third rule states that you can look for properties and classes found in data, as well as information from RDF, RDFS, and OWL ontologies, such as relationships between ontology items. To put it differently, to be able to use URIs efficiently, we should use RDF or SPARQL for querying. The RDF is a W3C-developed graph-based representation standard for data publishing and interchange on the Web. It's also utilized in semantic graph databases (also known as RDF triplestores), which are designed to store interconnected data and infer new facts from current ones.

The fourth rule is to connect the data we have into a web, where one may gain new insight into their data through the data links.(11)Links to other URIs, similar to the hypertext web, connect data and allow us to find different things. We maximize reuse and interlinking among existing data by connecting new information with existing resources, resulting in a densely interconnected network of machine-processable meaning.(12)

2.1.3 Linked Data vs Open Data:

Open data is data that can be freely used and disseminated by anybody.

However, open data is not the same as linked data. Without any ties to other data, Open Data can be made available to everyone. Data can be linked without being freely available for reuse and distribution at the same time.

An RDF database such as Ontotext's GraphDB is an example of LOD. It can manage large datasets from many sources and link them to Open Data, allowing for better knowledge discovery and data-driven analytics.(12)

2.1.4 Linked Open Data:

Linked Open Data (LOD) is a term used to refer to tools or platforms that support freely-connected (interlinked) resources or frameworks to allow for collection and integration of data (usually derived from various sources or formats) and provide useful information that can be accessed by machines or humans.(13) Linked Open Data is a powerful combination of Linked Data and Open Data in that it is both linked and based on open sources. DBpedia, a crowd-sourced community effort to extract structured information from Wikipedia and make it available on the Web, is an example of a LOD set.

A semantic graph database, like Ontotext's GraphDB, can handle large datasets from a variety of sources and connect them to Open Data. Richer queries result, allowing for more latent knowledge discovery and effective data-driven analytics.

To summarize, Linked Data takes information in distinct formats and dismantles the barriers that exist between those different sources. It promotes data model extension and allows for quick modifications. Data integration and navigating through complex data become considerably easier and more efficient as a result.

The connecting of heterogeneous sources and formats in semantic graph databases allows for the inference of new knowledge from existing facts. As a result, Linked Data enables people to place proprietary knowledge in the context of open-world knowledge and/or commercially specialized knowledge, hence enhancing cognitive and semantic technology innovation. (12).

2.1.5 Data Ethics:

Traditional Ethics:

According to MacIntyre et al. traditional deontological and utilitarian ethics have placed a major focus on individual moral responsibility, often known as moral agency, since the Enlightenment.(14). This concept of moral agency is based on nearly religiously held beliefs in individualism and free will. When it comes to the progress of modern technology, both of these assumptions face problems, particularly Open data. The degree to which an entity has moral agency determines its level of responsibility. The culpability of this entity is defined by moral responsibility in combination with extraneous and intrinsic variables that escape the creature's will. In general, various entity innate conditions determine moral agency, three of which are widely accepted, according to Noorman et al. (15)

1. **Causality:** An agent can be held responsible if the ethically relevant result is an outcome of its actions.

2. **Knowledge:** If an agent had (or should have had) knowledge of the repercussions of its acts, it can be held responsible for the outcome.

3. **Choice:** If an agent has the freedom to choose another option without causing more harm to itself, it could be held responsible for the outcome.

Observers implicitly prefer to excuse agents who lack full moral agency, i.e. when at least one of the three conditions is missing. However, there are lines of reasoning that analyze morally important outcomes without regard to the existence of a moral agent, at least in the sense that negative consequences impose moral obligations(16). Network ethics (17), social networking ethics (18), distributed and corporate moral responsibility (19), and computer and information ethics (20) have all recently produced significant breakthroughs in ethics. Nonetheless, Big Data & the Open data movement has brought forth other changes, such as the philosophical dilemma of 'many hands,' i.e. the influence of many individuals contributing to an action in the form of distributed morality (15), that must be addressed. Big data is a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications

(21).

When reviewing the key criteria of Big Data and Open data, it becomes evident that in some cases, data ethics moves away from personal moral agency. In other circumstances, it raises the moral culpability of people in charge of Big Data and Open data. However, there is a growing trend toward impersonal ethics focused on the repercussions for others(22). At the heart of data ethics are four relevant qualities:

[1.] There is more data now than there has ever been in data history (Smolan and Erwitt 2012):

- Beginning of recorded history till 2003—5 billion gigabytes had been created.
- 2010 — By this point, data volume of 2 zettabytes (2trillion gigabytes) created.
- 2015 — 15.5 zettabytes created
- 2021 — 79 zettabytes created.

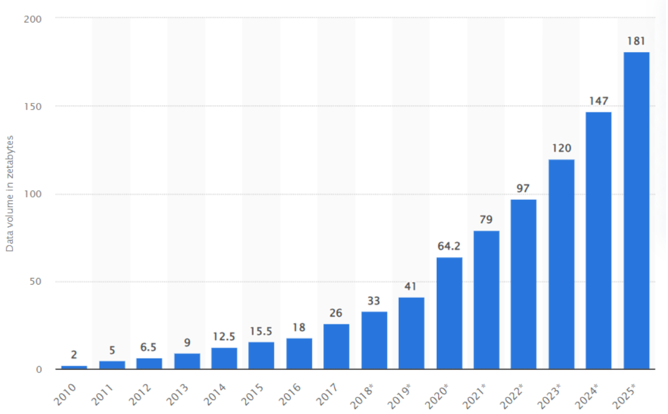


Figure 2.1: Volume of data created, captured, copied, and consumed worldwide from 2010 to 2025 (in zettabytes (One trillion gigabytes))

[2.] Open Data is organic: by gathering everything that is digitally available, it portrays reality digitally lot more naturally than statistical data it is much more organic in this respect. This messiness is the outcome of a representation of reality's messiness (for example, format discrepancies and measurement artifacts). We can get closer to a computer approximation of reality with it.

[3.] Big data & Open Data has the potential to be global: not only is the representation of reality organic, but with truly massive Big Data sets (such as Google's), the reach expands globally.

[4.] Correlations vs. causation: In data analytics, correlations take precedence over causation. (22)

There is a wide range of data that is directly or indirectly related to individuals and their interactions: social network data, the rising field of health tracking data, emails, text messaging, the simple usage of the Google search engine, and so on. Even while this type of data does not make up the all Data, it can be quite ethically problematic.

Data Ethics challenges:

The ethical difficulties become evident as you go through the four ethical features of Big Data listed above.

[1] and [2]: Just as global warming is the result of many individuals and organizations emitting greenhouse gases, Big Data & Open Data is the result of individual activities, sensory data, and other real-world measures resulting in a digital representation of our reality. Cukier calls this "datafication" (23). Already, the "data generator" (e.g., internet users, cellphone owners, person being vaccinated etc.) is at an ethical disadvantage regarding their knowledge and free will due to a lack of understanding about which data is collected or what it might be used for. The "internet of things" adds to the gap between one actor's knowledge and will and the source of information and power of the other.

[3]: Global data creates a power imbalance between many stakeholders, with corporate agencies benefiting the most from the essential know-how to generate insight and understanding from data.

[4]: Data correlations imply causation where none may exist. We grow more prone to believing what we perceive without understanding the underlying reasons. (22)

2.1.6 Data Ethics Legislation & GDPR:

The practice of defining and executing policies, procedures, and standards for the correct development, use, and control of the infosphere is known as digital governance. It's also an issue of convention and proper coordination, which isn't always moral or immoral, legal or criminal. A government agency or a company, for example, can use digital governance to (i) determine and control the processes and methods used by data stewards and data custodians in order to improve the data quality, reliability, access, security, and availability of its services; and (ii) devise effective procedures for decision-making and the identification of accountabilities with regard to data-related processes.

Digital governance may include principles and recommendations that are similar to digital regulation. This is just another way of referring to relevant legislation, which is a set of regulations devised and enforced by social or political institutions in order to regulate the behavior of relevant agents in the information sphere. Not every component of digital governance is a subject of digital regulation, and not every aspect of digital regulation is a matter of digital governance. The General Data Protection Regulation (GDPR) serves as a solid illustration in this case. The essential relationship through which digital legislation shapes digital governance is compliance.(24)

Digital regulation in the EU is now determined by the GDPR, and that EU legislation is normally respectful of human rights, it may be useful to understand the value of the distinction between soft and hard ethics and their relations to legislation by using the GDPR as a concrete case of application. This is the legislation that replaces the Data Protection Directive 95/46/EC. It is designed to harmonize data privacy laws across Europe, to protect and empower all EU citizens' data privacy, independently of geographical location, and to improve the way organizations across the EU approach data privacy. The GDPR comprises 99 Articles.(24) The Articles do not cover everything, leave grey areas of normative uncertainty even about topics they do cover, are subject to interpretations, and may require updating when applied to new circumstances, especially in a technological context where innovation develops so quickly and radically; consider face recognition software or so-called deep fake software, for example. It is obvious that the foresight analysis of the ethical impact of digital innovation, or simply ethical impact analysis (EIA), must become a priority. The task of digital ethics is not simply to 'look into the [digital] seeds of time, it also seeks to determine which ones should grow and which should not.(25)

Digital technologies have a wide range of benefits, but they also pose a number of challenges and risks. Assuring socially preferable outcomes necessitates resolving the tension between incorporating benefits and reducing potential drawbacks, or, in other words, promoting these technologies while avoiding their abuse, underuse, and destructive usage. This is when the importance of taking an ethical approach becomes clear. Digital ethics, in

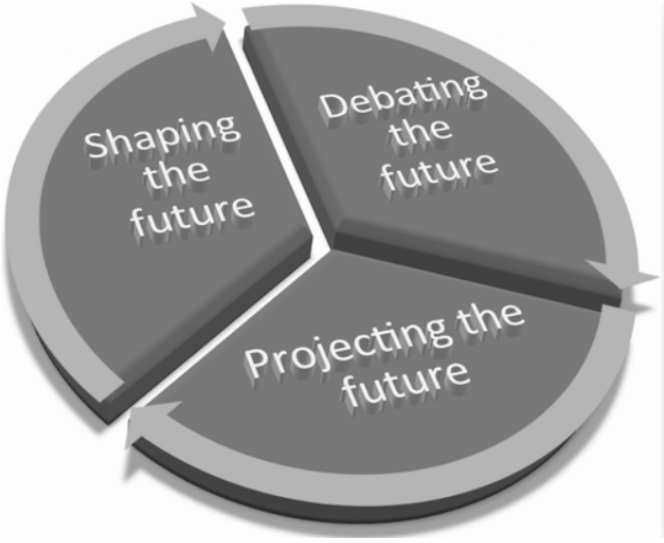


Figure 2.2: Ethical impact analysis (EIA) the foresight analysis cycle

particular, cannot be treated as an afterthought. It can't just be a questioning exercise. The development of critical awareness, as well as an ethical approach to the design and control of the digital world, are essential. Ethics must inform strategies for the development and use of digital technologies from the very beginning, when changing the course of action is easier and less costly, in terms of resources and impact. This is particularly relevant in the EU, as Floridi (24) has argued, where soft ethics can be properly exercised and where a soft ethical approach to SETI (science, engineering, technology and innovation) developments is acknowledged to be crucial. If soft digital ethics can be a priority anywhere, this is certainly in Europe.

A separate analysis by Hossain (26) analysis found that the current legal frameworks neither protect the interests of the associated parties nor encourage public data use & many other studies have emphasized the importance of developing a legal framework explaining publishing and using public data.

2.2 Technical Background:

2.2.1 RDF

RDF is a standard for data interchange that is used to describe data that is highly linked. Each RDF statement is a three-part structure made up of resources, each of which is designated by a unique URI. AI systems can readily identify, disambiguate, and connect information when it is represented in RDF.

RDF stands for Resource Description Framework, and it is a web resource and data interchange standard developed and standardized by the World Wide Web Consortium (W3C) (27). While there are many traditional methods for working with data and, more specifically, data connections, RDF is the most simple, powerful, and expressive standard currently available.

RDF uses triples (three positional statements) to link data fragments together.

In plain English, an RDF statement connects many types of resources to express facts, relationships, and data. Almost everything can be stated by a consistent structure consisting of three connected data items using an RDF statement. It is important to know that all data, regardless of its format, can be converted to RDF data.

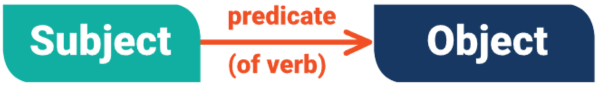


Figure 2.3: How triples work in English

Being a powerful and expressive framework for representing data, RDF is used for building knowledge graphs – richly interlinked, interoperable and flexible information structures. (1)

2.2.2 Knowledge Graphs

At the core of the knowledge graph lies a knowledge model — a set of interconnected descriptions of concepts, entities, relationships, and events.

Descriptions contain formal semantics that allow humans and computers to process them efficiently and without ambiguity.

Descriptions feed into one other, establishing a network in which each entity represents a piece of the description of the entities that are associated to it.

According to the knowledge model, diverse data is related and characterized by semantic metadata.

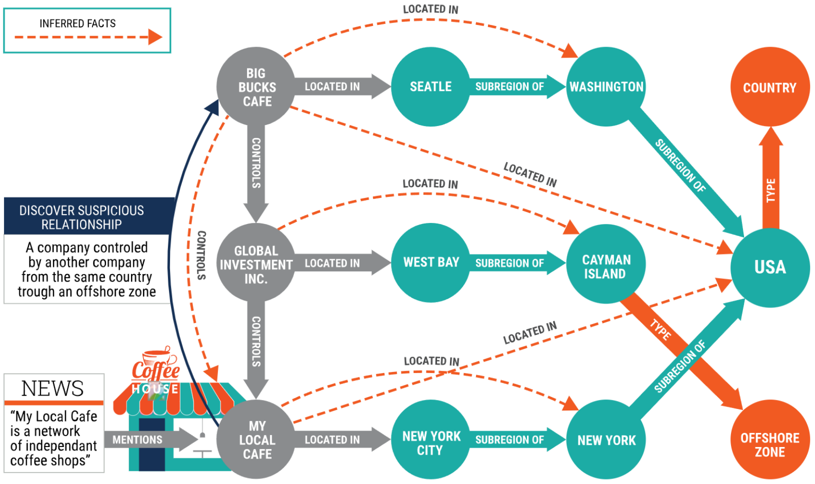


Figure 2.4: Example of how Knowledge graphs interlink data

Knowledge graphs integrate features from a variety of data management concepts:

- Databases, because the data can be explored via structured queries

-Graphs, because they can be analyzed as any other network data structure;

- Knowledge bases, because they bear formal semantics,

which can be used to interpret the data and infer new facts.

Knowledge graphs, which are represented in RDF, include the following features which give the ideal framework for data integration, unification, linking, and reuse:

Expressivity: The Semantic Web stack's standards – RDF(S) and OWL – enable for the fluent representation of a wide range of data and content, including data schema, taxonomies and vocabularies, numerous types of metadata, reference and master data.

Performance: All of the specifications have been carefully considered and tested to ensure efficient administration of graphs containing billions of facts and characteristics.

Interoperability: Data serialization, access, administration, and federation are all covered by a variety of specifications. Data integration and publication are made easier with the usage of globally unique identifiers (URIs).

Standardization: All of the above is standardized through the W3C community process to ensure that the needs of many actors – from logicians to enterprise data management specialists and system operations teams – are met. (1)

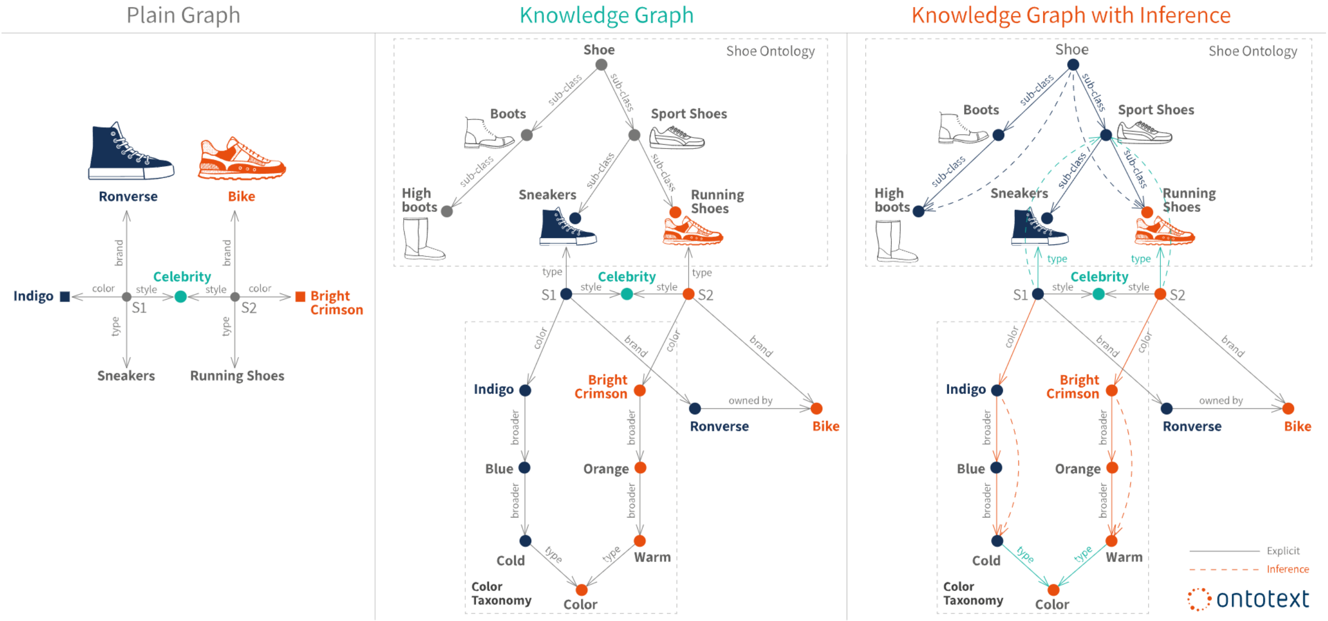


Figure 2.5: Plain graph, Knowledge Graphs & Knowledge graph with Inference (1)

2.2.3 Ontologies:

Ontologies are the foundation of a knowledge graph's formal semantics. Semantics describe the processes a computer follows when executing a program in that specific language. Ontologies can be thought of as the graph's data schema or formalised plan of how the graph is organised. They serve as a formal contract between the knowledge graph's creators and its users over the meaning of the data included inside it. A user could be another person or a computer program that needs to interpret facts in a reliable and exact manner. Ontologies ensure a shared understanding of the data and its meanings.

The ontology data model can be applied to a set of individual facts to create a knowledge graph – a collection of entities, where the types and the relationships between them are expressed by nodes and edges between these nodes, By describing the structure of the knowledge in a domain, the ontology sets the stage for the knowledge graph to capture the data in it.

There are a number of representation and modeling tools that can be utilized when formal semantics is used to express and evaluate the data of a knowledge graph:

Classes An entity description will almost always include a classification of the entity in terms of a class hierarchy. When dealing with business data, for example, there could be classes such as Person, Organization, and Location. A common superclass Agent can be shared by persons and organizations. Country, populated place, city, and so on are all sub-classes of location. Object-oriented design borrows the concept of class, with each entity usually belonging to only one class.

Relationship Types Relationships between entities are typically marked with types, which provide information about the relationship's nature, such as friend, relative, competitor, and so on. Relationship types can also have formal definitions, e.g., that parent-of is inverse relation of child-of, they both are special cases of relative-of, which is a symmetric relationship. Alternatively, sub-region and subsidiary can be defined as transitive relationships.

Categories Categories that describe some feature of an entity's semantics, such as "18th century composers," might be connected with it. A book can fall into multiple categories at the same time, such as "Books about travel", "Bestsellers", "Books by Italian authors" and so on. The categories are described and ordered into taxonomy.

Ontologies are part of the W3C standards stack for the Semantic Web as one of the building components of Semantic Technology. They give users the structure they need to connect one piece of data to other pieces of data on the Web of Linked Data. Ontologies facilitate database interoperability, cross-database search, and smooth knowledge management by specifying common modeling representations of data from remote and heterogeneous systems and databases.

Knowledge graphs are used in a variety of ways and for a variety of purposes. Intelligent content and package reuse, responsive and contextually aware content recommendation, semantic search, investment market intelligence, information discovery in regulatory documents, advanced drug safety analytics, and other data and information-intensive services are examples. (1)

2.2.4 SPARQL

SPARQL is the standard query language and protocol for Linked Open Data and RDF databases. Users can query information from databases or any data source that can be mapped to RDF using SPARQL, which stands for "SPARQL Protocol and RDF Query Language." The W3C established and endorsed the SPARQL standard, which allows users and developers to focus on what they want to know rather than how a database is organized. A SPARQL query is made up of a series of triple patterns, in which each element (the subject, predicate and object) can be a variable. The variables' solutions are then discovered by comparing the patterns in the query to triples in the dataset.

There are four different types of queries in SPARQL. It can be used for a variety of purposes, including:

ASK if the RDF graph data contains at least one match for the query pattern;

SELECT all or some of the matches in a tabular format (including aggregate, sampling, and pagination);

CONSTRUCT an RDF graph by substituting the variables in these matches in a set of triple templates;

DESCRIBE the matches that were discovered by creating a suitable RDF graph.

SPARQL can efficiently extract information hidden in non-uniform data and stored in various formats and sources.

2.3 State of the Art

This section covers a study of Linked Open data tools currently freely available and accessible.

2.3.1 R2RML Tools

Some tools in the area of R2RML mappings include Oracle Spatial and Graph 19c includes an open, scalable, secure and reliable RDF management platform. Based on a graph data model, RDF triples are persisted, indexed and queried, similar to other object-relational data types. Oracle Spatial is used to create interactive maps and perform spatial analysis on business data quickly and easily. Users can visualize, explore, and analyze geospatial data stored in and managed by Oracle in the cloud or on-premises.

Another similar tool is OpenLink Virtuoso. It is a SQL-ORDBMS and Web Application Server hybrid (aka Universal Sever) that provides SQL, XML, and RDF data management. Triple Store access is available via SPARQL, SIMILE Semantic Bank API, ODBC, GRDDL, JDBC, ADO.NET, XMLA, WebDAV, and Virtuoso/PL (SQL Stored Procedure Language). The product is available in Open Source and Commercial editions. (28)

2.3.2 Ontology Design tools

To design the ontology used in this project, an Ontology design tool was needed. There are many such tools available such as: KAON, Protégé, WebODE, OntoEdit, SWOOP, HCONE, SOBOLEO, ORIENT, NeOn Toolkit, TopBraid Composer, OBO-Edit, and DIP Ontology Management Suite.

Protégé (Noy et al., 2000) was developed by Stanford Medical Informatics at the Stanford University School of Medicine. It is used in many areas where the concepts can be modeled as a class hierarchy. It is a Java based open-source tool for editing and managing ontologies. It is the most widely used domain-independent, freely available, platform-independent technology for developing and managing terminologies, ontologies, and knowledge bases in a broad range of application domains. With Protégé it is easy to create classes and hierarchies, to declare properties for classes, create instances and introduce values, all these under an environment consisting in menus, buttons, dialog boxes and easy to use graphic representations. (29)

2.3.3 RDF Triple store

An RDF triple store was needed to hold the RDF triples from the R2RML mapping & Uplift. There are many available open source tools to do so. Some of these include: Oracle Spatial and Graph with Oracle Database, AnzoGraph DB by Cambridge Semantics, AllegroGraph, OpenLink Virtuoso, Stardog, RDFox, & GraphDB™ by Ontotext. (30)

All of these were good options for this project, however I had some previous experience with Graph DB, therefore it was chosen to be most suitable for this project.

2.3.4 User Interface tools

There were many tools for design of a User Interface to run the competency question queries. The functionality required was to be integrated with SPARQL, and suitable to develop an interesting interface. Some suitable tools were Streamlit, Dash, Shiny, Voila, & Flask. Streamlit, Dash, and Panel are full dash-boarding solutions, focused on Python-based data analytics and running on the Tornado and Flask web frameworks. Shiny is a full dash-boarding solution focused on data analytics. Voila is a library that turns individual Jupyter notebooks into interactive web pages. Flask is a Python web framework for building websites and apps from the ground up.

2.4 Overall Analysis of State of Art

From my analysis of the state of the art in tools to develop a Linked open data application, there were many different tools available freely online. The overall process for choosing which tools to use was using tools which had good online support, and tools which I had previous experience using.

3 Design

This chapter outlines the requirements identified to meet research objectives, including functional and non-functional requirements, and then functional architecture, design challenges and the corresponding design decisions are described.

3.1 High Level Requirements

The research objectives as discussed in section 1.3 suggests the following need to be catered for in the design.

Requirement 1. Obtaining datasets which could potentially be linked together in what could be considered an ethically questionable manner.

Objective: Finding several datasets which contain factors which can be linked together.

Approach: Searching the internet for various datasets with commonalities to be linked together.

Requirement 2. Develop a set of requirements which to assess what fits as an Ethically questionable area.

Objective: Setting out what makes a dataset have the potential to be "Ethically ambiguous"

Approach: Rely on research on ethics to determine a basic set of rules.

Requirement 3. Be able to assess information in the datasets and categorise them to highlight the potential risk areas.

Objective: A method is needed to produce a basic report about the ontology regarding the vocabulary prefixes, Type of data based on the type of vocabulary used, GDPR areas affected.

Approach: Develop an Ontology from the metadata and the set of competency questions that can be asked of a dataset.

Requirement 4. Developing a tool which would display the some "Competency questions" online to query the linked database.

Objective: A platform is needed to display the "Competency questions" and run queries on the database.

Approach: Looking into various tools available and using them to develop the User Interface.

3.1.1 Functional Requirements:

To address the high-level requirements, the following functional requirements were derived to achieve the project goals:

1. Find the a collection of datasets that fulfil requirements for being potentially ethically questionable.

2. Find a way to categorise datasets & find similar features in datasets.

3. Gather input data in form of the metadata, dataset information etc.

4. Clean datasets to ensure accuracy.

5. Create an ontology to define relationships between classes.

6. Define an R2RML mapping based on the ontology to generate RDF triples from datasets.

6. Create a knowledge base to store the above-generated information, and visualise it.

7. Create an interface to query the knowledge base.

3.2 Use Case

The steps above were followed to form the use case which was realized to test the principles of data ethics and investigate potential public data misuse. The User Interface application allows the user to do the following:

1. View Data sets - The user can browse the datasets and view the individual data entities in them, as well as their URIs (RDF triple Data labels).

2. Visualize Knowledge base - The User can view the datasets in a knowledge graph and view connections between data entities.

3. Query knowledge graph - The user can select from a number of SPARQL queries to run on the knowledge base.

4. Gain new insight into datasets - Using the above steps the User can gain new knowledge from the datasets as a result of the data linking.

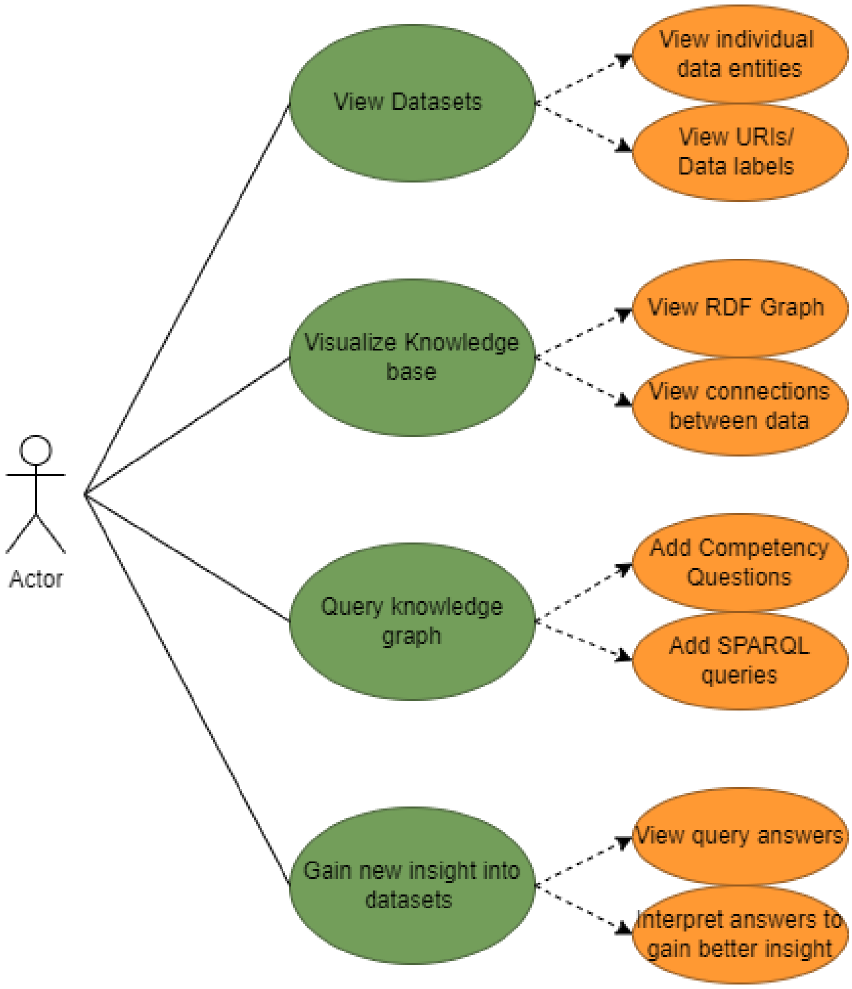


Figure 3.1: Use case of Linked open data application

3.3 Functional Architecture

The system's functional architecture is depicted in the diagram below. A basic description of the system's operation is provided below. The following section contains a more extensive explanation.

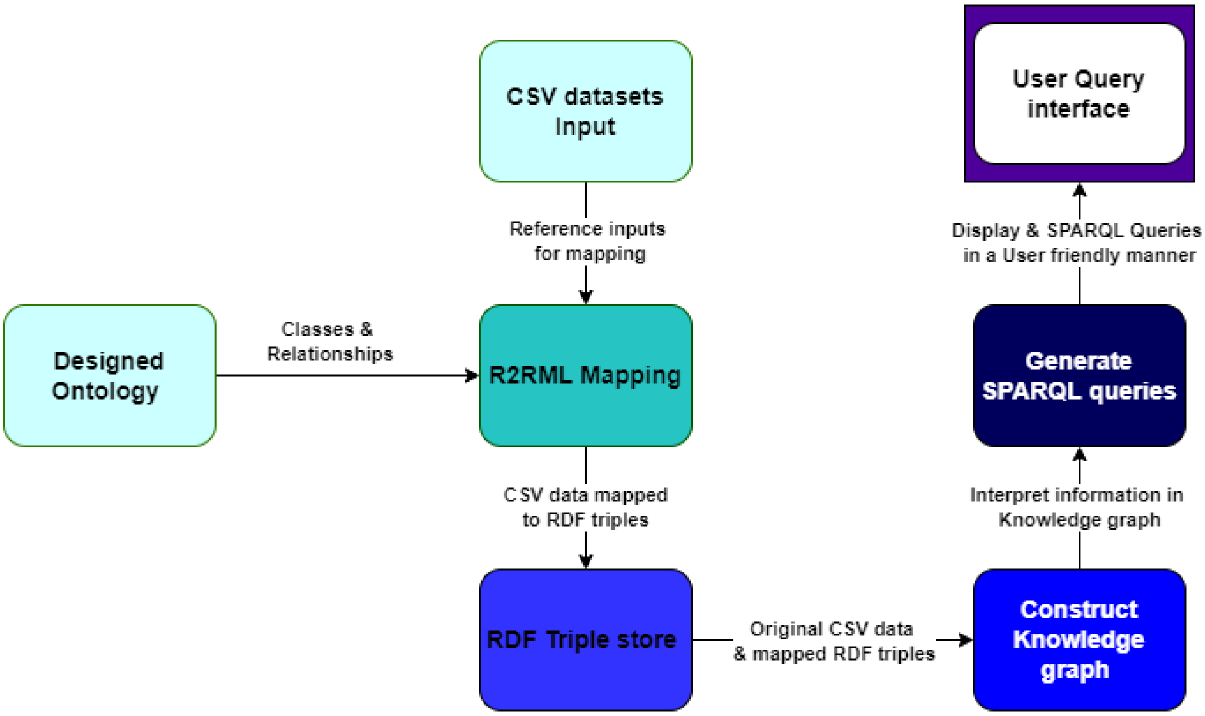

CSV datasets Input
User Query interface
Reference inputs for mapping
Display & SPARQL Queries in a User friendly manner
Designed Ontology
Classes & Relationships
R2RML Mapping
Generate SPARQL queries
CSV data mapped to RDF triples
Interpret information in Knowledge graph
RDF Triple store
Original CSV data & mapped RDF triples
Construct Knowledge graph

Figure 3.2: Functional Architecture of Linked open data application

The functional component description is as follows:

1. CSV Datasets input. This unit is responsible for providing the data which is going to be mapped. Reference inputs (A sample data from each column) are fed into the R2RML mapping.

2. Designed Ontology. This unit shows that the R2RML mapping is based on the classes and relationships from the designed ontology.

3. R2RML Mapping. R2RML mappings are used for expressing customized mappings from relational databases to RDF datasets. In this unit the CSV data is transformed into RDF triples.

4. RDF Triple store. This component signifies the RDF triples which are organised to be fed into the Knowledge graph.

5. Construct Knowledge graph. This component shows the Knowledge graph being setup (in Graph DB). The CSV data and the mapping.ttl file are used to do so.

6. Generate SPARQL queries. This component is responsible for setting up the SPARQL queries or compentency questions to demonstrate the data links present.

7. User Query interface. This component is used to run &display the SPARQL queries in a user friendly manner.

3.4 Design challenges & How Addressed

1. How to find suitable datasets ?

This challenge was addressed by researching the available datasets for download on various government's websites. It was important to find datasets which had commonalities or points which could be linked together. At the same time it was important that these datasets had a suitable ethical aspect to them, i.e they had to relate to everyday peoples lives, and could have a potential effect on them.

2. How to Develop a suitable Ontology ?

When the datasets had been found, it was important to develop a concrete ontology to link them together and establish clear classes which could be explored further.

3. How to generate suitable competency questions ?

To generate suitable competency questions it was important to find aspects of the data which could be looked at in a potentially ethically questionable way. Some aspects included areas with low vs high vaccination rates & the differences between richer and poorer areas.

3.5 Summary

In this chapter, the experimentation's design has been established. This design was created for the development of the Linked open data application, which was based on the research goals and specifications described in Chapter 2 - State of the Art Review.

4 Implementation

This chapter includes the programming infrastructure used to build the Linked open data application, the implementation and architecture of the different phases of the application, the ontology and the competency questions which have been used to derive the SPARQL queries.

4.1 Programming Infrastructure

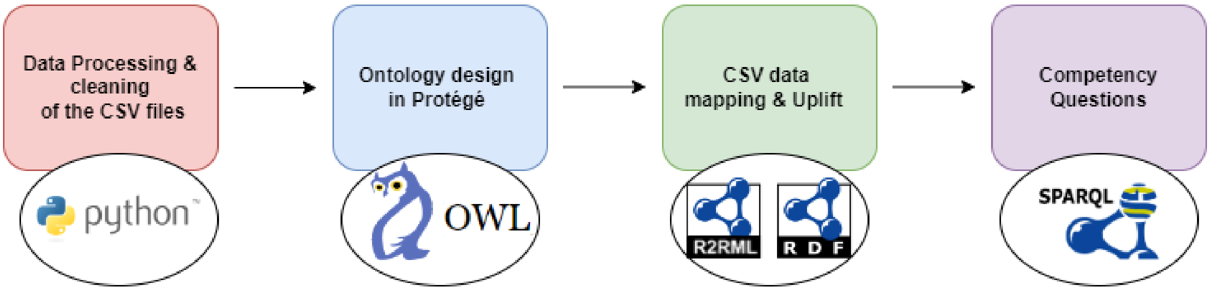

Data Processing & cleaning of the CSV files
python
Ontology design in Protégé
OWL
CSV data mapping & Uplift
R2RML
R D F
Competency Questions
SPARQL

Figure 4.1: Design Phases of Linked open data application

To start with, we needed to process the datasets and perform some basic data cleaning before linking it. Data cleaning of the CSV files was carried out using a simple Python script. The CSV files were imported using the Python library Pandas. They were loaded into matrices with column headings.

4.1.1 Data &CSV files

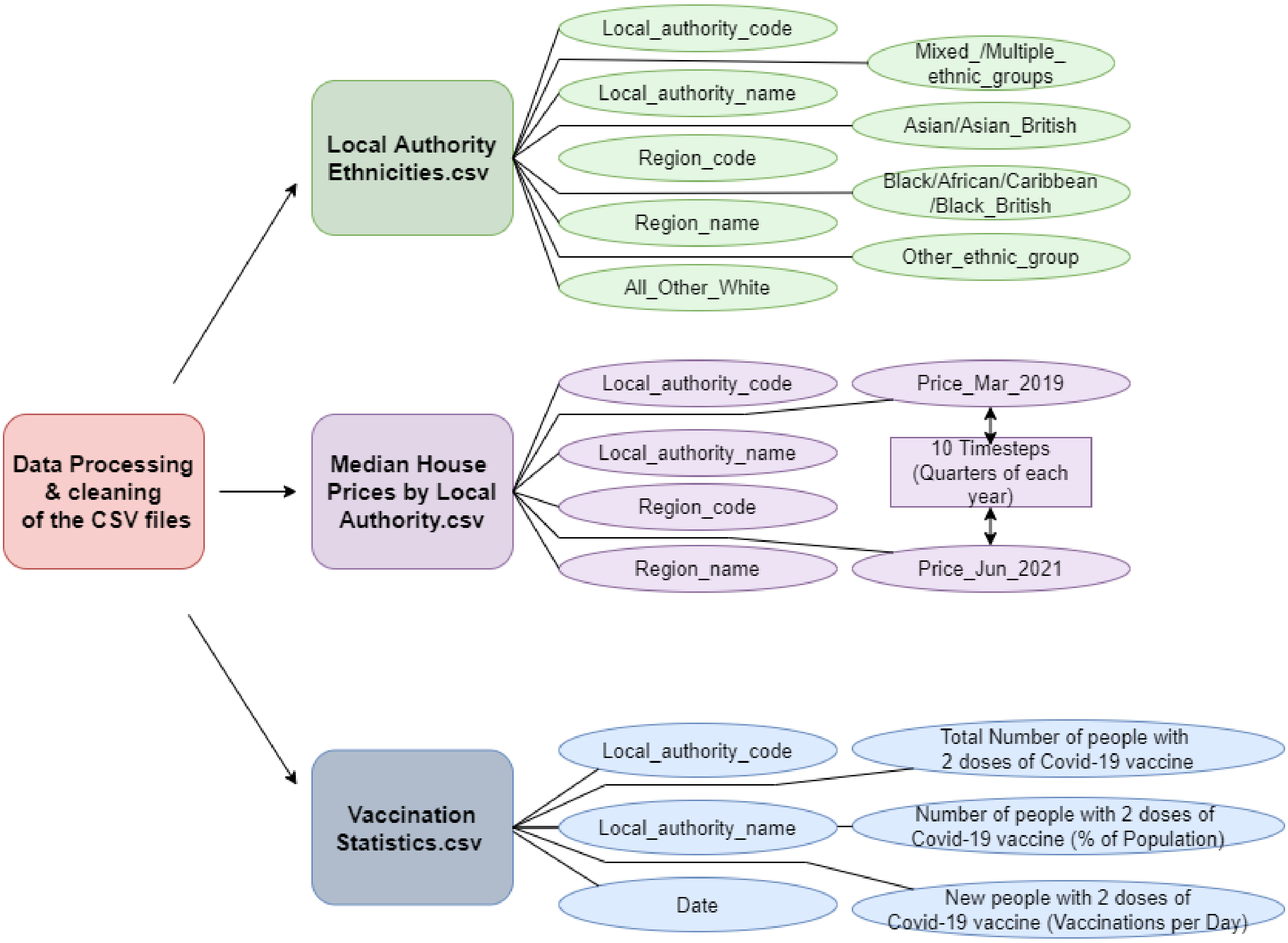

Data Processing & cleaning of the CSV files
Local Authority Ethnicities.csv
Local_authority_code
Mixed_/Multiple_ ethnic_groups
Local_authority_name
Asian/Asian_British
Region_code
Black/African/Caribbean /Black_British
Region_name
Other_ethnic_group
All_Other_White
Median House Prices by Local Authority.csv
Local_authority_code
Price_Mar_2019
Local_authority_name
10 Timesteps (Quarters of each year)
Region_code
Region_name
Price_Jun_2021
Vaccination Statistics.csv
Local_authority_code
Total Number of people with 2 doses of Covid-19 vaccine
Local_authority_name
Number of people with 2 doses of Covid-19 vaccine (% of Population)
Date
New people with 2 doses of Covid-19 vaccine (Vaccinations per Day)

Figure 4.2: CSV Files and their Collumn Headings

As we can see from the figure above, three CSV data files were linked in this project. "*Local Authority Ethnicity.csv*" contains information on the different ethnic groups that live in each "Local Authority area" in the United Kingdom. A local Authority area is one of 333 individual jurisdictions in the UK made up of councillors who are elected by the public in local elections (31). The second CSV file "*Median House Prices by Local Authority.csv*" includes data on Median House Prices in each Local Authority over 10 time steps from March 2019 to Jun 2021. Finally Vaccination Statistics provides data on the total number of people who received two doses of Covid-19 vaccine in each local authority from December 30, 2021, to December 8, 2020 (383 days).

There are 121 different local authorities in each CSV file. This is the main linkage area of commonality between the different CSV files. They are also linked by dates.

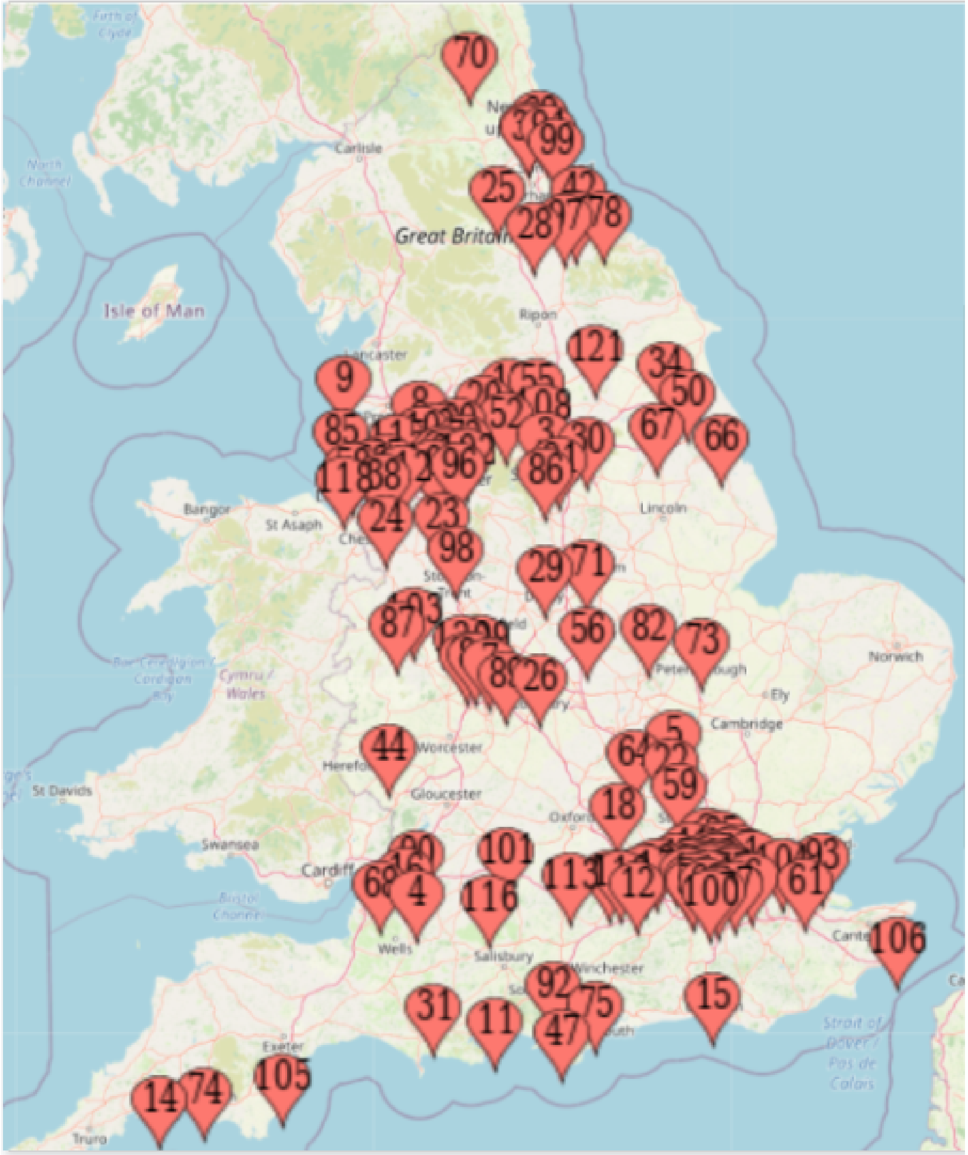


Figure 4.3: Map of the different Local Authority regions in the CSV files

The first dataset *Local Authority Ethnicities.csv* contains total populations (in thousands) in each local authority region, for each Ethnicity as given in figure 4.2 above. The dataset was published on 04 December 2019 by the Office for National Statistics in the Population characteristics research tables.

The second dataset *medianHousePrices.csv* contains median house prices for 10 time steps over a 3 year period for each Local authority. Also gives Region code and Region name to enrich the other datasets. Median house prices were used as they are a better indicator of the average house price, other measures like the mean are affected by outliers which can be common in the housing market. The source of this dataset was Median house prices for administrative geographies: HPSSA dataset 9 - Office for National Statistics.

Finally the last dataset *Vaccination Statistics.csv* provides data on the total number of people who received two doses of Covid-19 vaccine. The data spans from December 30, 2021, to December 8, 2020 (383 days). This dataset contains 46870 rows, each row corresponding to a given date in the time span and an individual local authority. The source of this dataset was https://coronavirus.data.gov.uk/.

These datasets were chosen to try form a concrete Use case and to demonstrate some ethical challenges

4.1.2 Ontology Design

Secondly the ontology was designed in Protege and was expressed using OWL (Web Ontology Language),

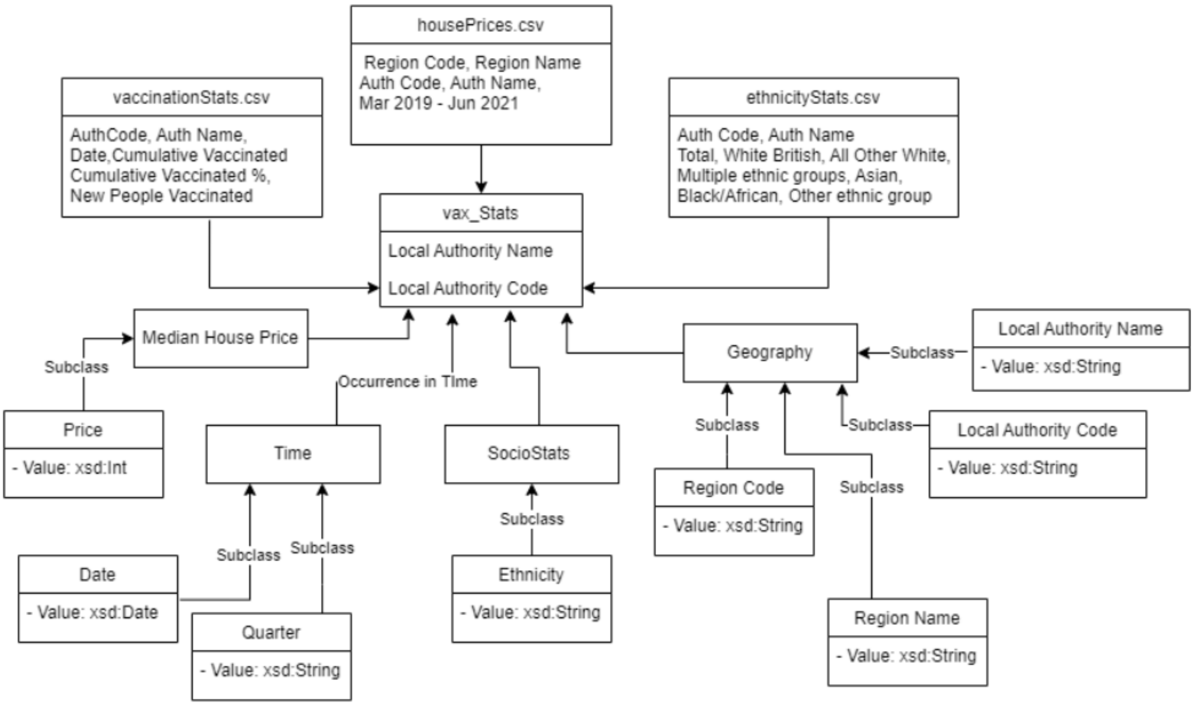


Figure 4.4: Ontology for Data linking application

The ontology above was designed with a purpose of showcasing the links and commonalities between the different CSV files discussed at length above. The classes the CSV files were linked to were House price, Time, SocioStats, and Geography. The ontology gives a schema on which to base the following R2RML mapping and Uplift section.

4.1.3 Mapping & Uplift

Next the CSV data was mapped according to the ontology using the R2RML mapping language.

R2RML is a language that allows you to define custom mappings from relational databases to RDF datasets. Existing relational data in the RDF data model can be seen using these mappings, which are described in the mapping author's preferred structure and target language, in this case the above ontology. Every R2RML mapping is custom-made for a specific database schema and vocabulary. A relational database that follows that schema is used as the input to an R2RML mapping. The result is an RDF dataset that employs predicates and types from the target vocabulary, as described by SPARQL. R2RML mappings are authored in Turtle syntax and are expressed as RDF graphs.

For this project a custom R2RML mapping was authored to convert each CSV file to a set of RDF triples. The mapping gives a template for each column heading in each CSV file is to be interpreted. The language is written in triples which are processed when the data is being mapped. An example of how to map the Median House price csv to the Geography class is given below.

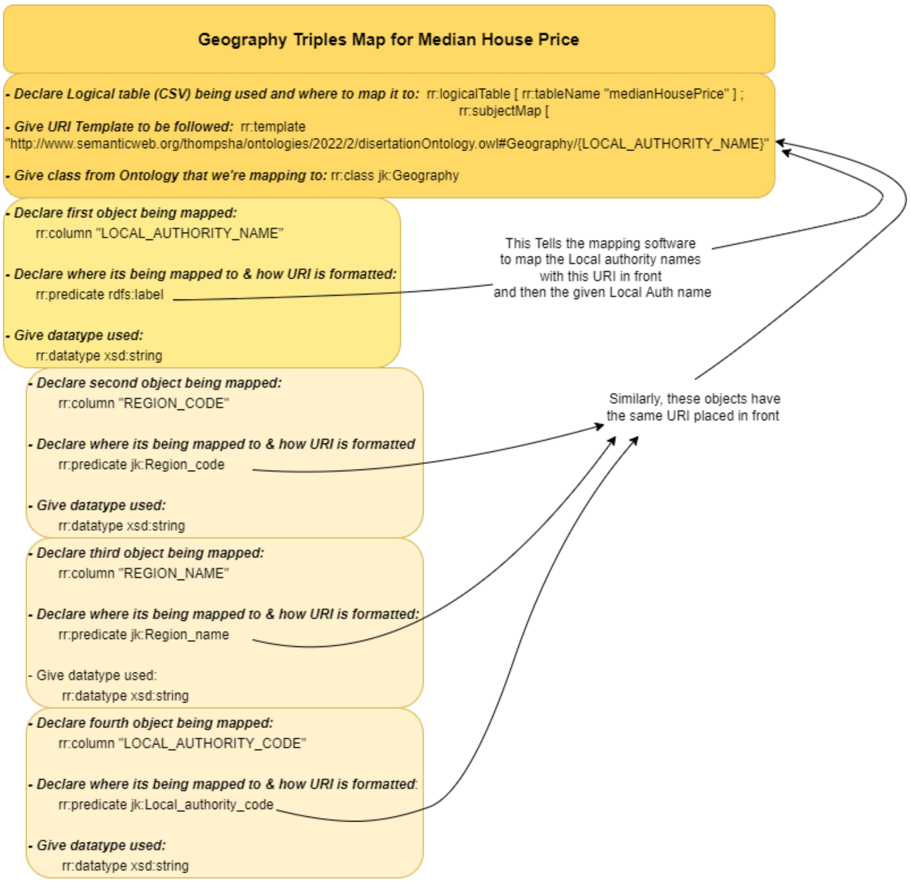

Geography Triples Map for Median House Price
- Declare Logical table (CSV) being used and where to map it to: rr:logicalTable [rr:tableName "medianHousePrice"] ;
rr:subjectMap [
- Give URI Template to be followed: rr:template
"http://www.semanticweb.org/thompsha/ontologies/2022/2/disertationOntology.owl#Geography/{LOCAL_AUTHORITY_NAME}"
- Give class from Ontology that we're mapping to: rr:class jk:Geography
- Declare first object being mapped:
rr:column "LOCAL_AUTHORITY_NAME"
- Declare where its being mapped to & how URI is formatted:
rr:predicate rdfs:label
- Give datatype used:
rr:datatype xsd:string
This Tells the mapping software
to map the Local authority names
with this URI in front
and then the given Local Auth name
- Declare second object being mapped:
rr:column "REGION_CODE"
- Declare where its being mapped to & how URI is formatted
rr:predicate jk:Region_code
- Give datatype used:
rr:datatype xsd:string
Similarly, these objects have
the same URI placed in front
- Declare third object being mapped:
rr:column "REGION_NAME"
- Declare where its being mapped to & how URI is formatted:
rr:predicate jk:Region_name
- Give datatype used:
rr:datatype xsd:string
- Declare fourth object being mapped:
rr:column "LOCAL_AUTHORITY_CODE"
- Declare where its being mapped to & how URI is formatted:
rr:predicate jk:Local_authority_code
- Give datatype used:
rr:datatype xsd:string

Figure 4.5: Mapping Median House price to Geography class

As we can see in figure 4.5, some specific data from the CSV file Median house prices is mapped, in accordance with the ontology, to RDF triples. Each column is taken in, the URI is added and the datatype of the data is assigned. This process is followed for all of the various classes of the ontology and all of the datasets.

These RDF triples were then uploaded to the triplestore Graph DB to visualise the data and form the SPARQL queries.

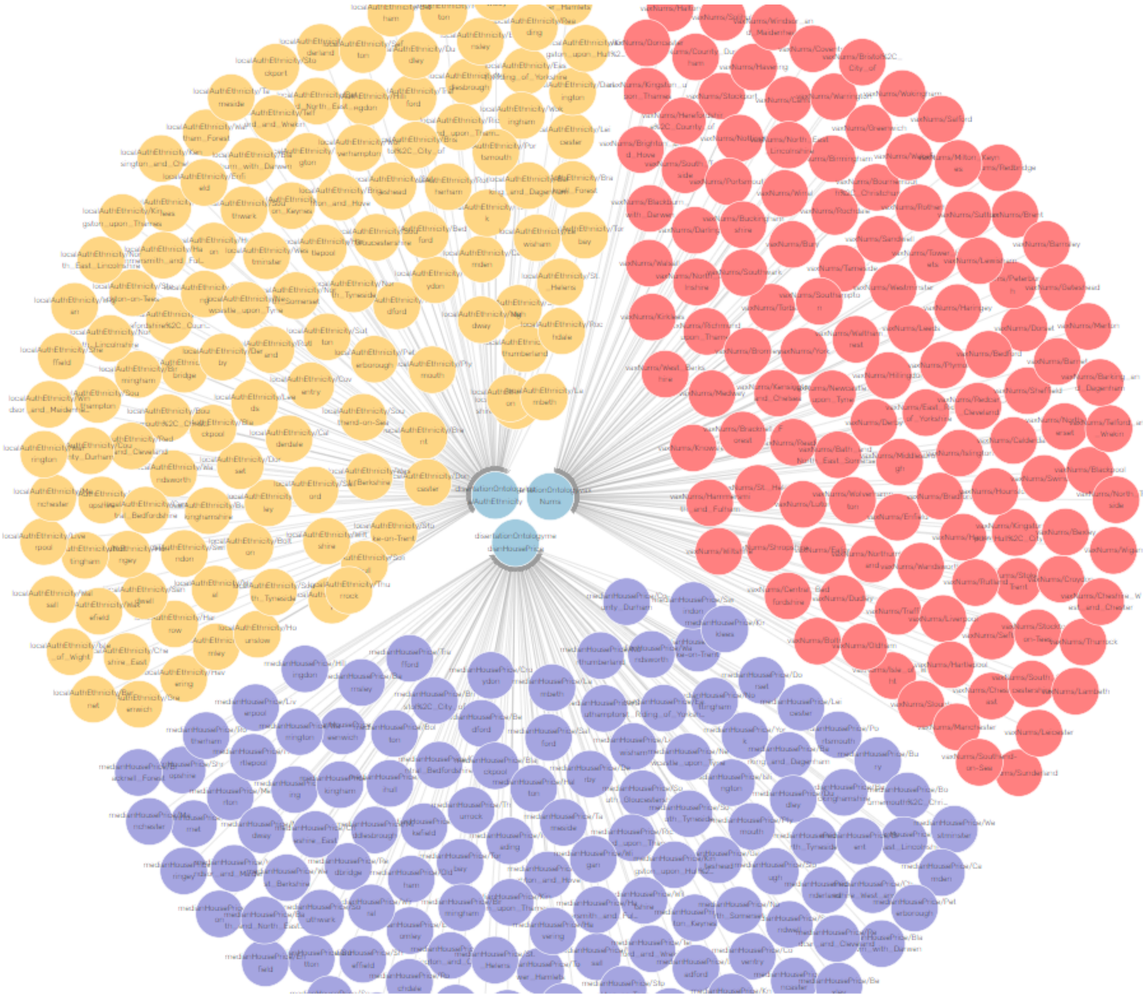

dissertationOntology:localAuthEthnicity
dissertationOntology:vaxNums
dissertationOntology:medianHousePrice

Figure 4.6: RDF Triples in Knowledge graph

We can see above the data from the three datasets uploaded to the triple store. The Ethnicity data is shown in yellow, the Median house price data is shown in purple, and the vaccination statistics shown in red. The benefit of this is that we can now utilise SPARQL to investigate the potential different links between the datasets, and then analyse the potential ethic impact.

4.1.4 Competency Questions

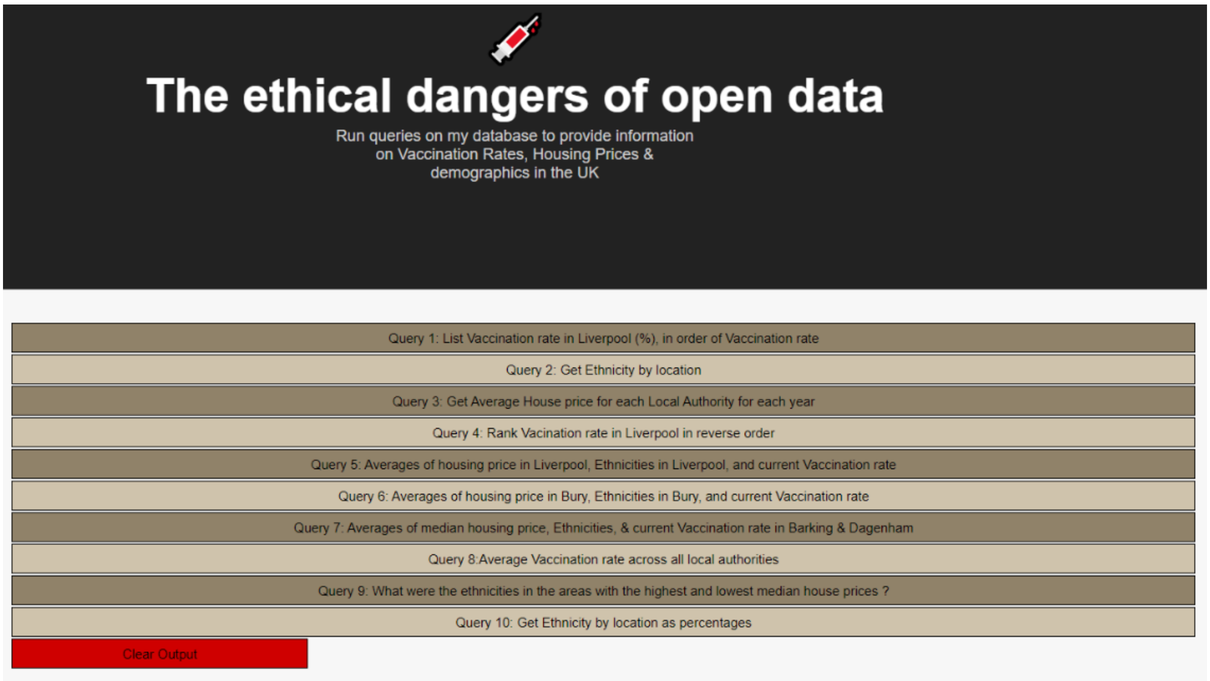
Finally the competency questions were expressed using SPARQL. These are

List of Queries	
Query 1:	List Vaccination rates in Liverpool, in order of Date
Query 2:	Get Ethnicity by Local Authority
Query 3:	Get Average House price for each Local Authority for each year
Query 4:	List Vaccination rates in Liverpool by date, in reverse order
Query 5:	Averages of median housing price , Ethnicities, & current Vaccination rate in Liverpool
Query 6:	Averages of median housing price, Ethnicities, & current Vaccination rate in Bury
Query 7:	Averages of median housing price, Ethnicities, & current Vaccination rate in Barking & Dagenham
Query 8:	What is the average Vaccination rate across all local authorities ?
Query 9:	What were the ethnicities in the areas with the highest and lowest median house prices ?
Query 10:	Get Ethnicity by location as percentages

Table 4.1: List of Competency Questions/Queries

These competency questions were designed to investigate the various datasets and the links between them. This final design phase serves to highlight the fact that anyone could publish this data in a publicly accessible format. They were also developed in line with the Design challenges addressed in section 3.4.

These questions are displayed in a User Interface:



The ethical dangers of open data

Run queries on my database to provide information
on Vaccination Rates, Housing Prices &
demographics in the UK

Query 1: List Vaccination rate in Liverpool (%), in order of Vaccination rate
Query 2: Get Ethnicity by location
Query 3: Get Average House price for each Local Authority for each year
Query 4: Rank Vacination rate in Liverpool in reverse order
Query 5: Averages of housing price in Liverpool, Ethnicities in Liverpool, and current Vaccination rate
Query 6: Averages of housing price in Bury, Ethnicities in Bury, and current Vaccination rate
Query 7: Averages of median housing price, Ethnicities, & current Vaccination rate in Barking & Dagenham
Query 8:Average Vaccination rate across all local authorities
Query 9: What were the ethnicities in the areas with the highest and lowest median house prices ?
Query 10: Get Ethnicity by location as percentages
Clear Output

Figure 4.7: User Interface

4.2 Technologies used:

This section details the various technologies, programs and applications used to create the Linked open data application.

The main libraries and APIs that were used to do so were:

1. NumPy & Pandas are both python libraries used for the data cleaning. NumPy a python library is used for creating and handling large arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays. Pandas is a python library for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series.

2. Dash is a python library which is used for creating interactive web applications. Dash is used to build analytical web applications without requiring advanced web development knowledge. In this project it was used to construct the User Interface component where the SPARQL queries were run.

3. SPARQLWrapper was used to process the SPARQL queries in the User Interface. SPARQLWrapper is a simple Python wrapper around a SPARQL service to remotely execute queries. It helps in creating the query invocation and, possibly, convert the result into a more manageable format.

4.2.1 R2RML

As discussed previously, R2RML is a programming language for creating bespoke mappings between relational databases and RDF datasets. These mappings, which are given in the mapping author's preferred structure and target language, in this case the aforementioned ontology, can be used to see existing relational data in the RDF data format. Each R2RML mapping is created specifically for a database schema and vocabulary. The input to an R2RML mapping is a relational database that follows that schema. As a result, an RDF dataset with predicates and types from the target vocabulary has been created, as described by SPARQL. R2RML mappings are expressed as RDF graphs and are written in Turtle syntax.

This R2RML mapping was processed by a purpose built mapping written in Java, which used many different java based dependencies such as SLF4J (Simple Logging Facade for Java) & Maven dependencies. This used a CSV configuration file to declare the CSV files, the mapping file and the output files. The template from the mapping file was used to go through each CSV to map their triples to the output file which was saved in Turtle syntax, and uploaded to GraphDB.

4.2.2 Protégé

Protégé is an ontology editor and knowledge management system that is free and open source. To define ontologies, Protégé provides a graphical user interface. Deductive classifiers are also included to ensure that models are consistent and to infer new information from an ontology's investigation. This application is written in the Java programming language.

In this project, Protégé was used to produce the ontology, which the R2RML mapping was based on. The various classes and relationships between them were input, and the software infers connections between classes. This helped give a schema to follow when developing the mapping.

4.2.3 GraphDB

GraphDB is an RDF database that is fast, reliable, and scalable. It makes loading and using linked data datasets much easier. GraphDB supports all RDF serialization formats and implements the RDF4J framework APIs as well as the W3C SPARQL Protocol specification. GraphDB is one of the few triplestores that can do semantic inference at scale, allowing users to create new semantic facts from old ones. Massive loads, queries, and inferencing are all handled in real time.

GraphDB was used in this project as the RDF triplestore, where the RDF triples produced in the R2RML mapping were uploaded. The Knowledge graph was then created from this triple store, which was helpful in visualising the datasets.

Finally the SPARQL queries were written in GraphDB according to the competency questions. There is a built in SPARQL editor in GraphDB, and the various queries were tested and developed using this workbench.

4.2.4 Dash

Dash is an open source framework for creating data visualization interfaces that may be used to construct dynamic online apps without learning Javascript or Front End Web development. It is a python library.

It was used to run and display the Competency questions in a simple, User friendly format. The queries were executed in the Dash environment which was run through Jupyter Notebook. First the Dash various elements which would be needed to display the data were imported, as well as some matrix processing libraries like Pandas to handle the output from GRaphDB, and also a SPARQL Wrapper to handle the SPARQL queries and send them to GraphDB to be executed. The various buttons and styling were also. Using Dash each query was sent to the GraphDB repository, and the result from each query was briefly processed to improve things like column headings, converting to percentages etc.

4.3 Summary

In this chapter, it has been discussed about the programming setup that had been used for implementing the Linked Open Data application, the architectural details and making of the of the different phases of the application. It also covered the design of the ontology, mapping and the competency questions to query the database.

5 Result & Evaluation

This chapter covers the motivation for the experiment, experiment goal and hypothesis, experimental setup, experiment results and analysis.

5.1 Motivation

The Linked Open data application has been developed as a prototype tool to perform the basic functionality:

To outline and highlight ethical & privacy concerns involved with the large scale release of public/open data. The Linked Open data tool provides a query interface to answer questions based on the linked datasets, and the links between them. The output of the application needs to be validated & analysed to review its effectiveness and usefulness in highlighting ethical concerns.

5.2 Experiment Goal & Hypothesis

The goal of the experiment is that the linked Open data tool must be able to answer a set of competency questions and demonstrate some links between datasets .

To evaluate the goal, following Hypothesis have been formulated:

1. The prototype application can answer competency questions correctly.
2. The competency questions demonstrate links between data.
3. The linking of the data presents some potential questions about ethics & dangers to the public.

These Hypotheses will be tested to explore whether the project goals have been achieved.

5.3 Experimental Setup

The experiment containing the input datasets, list of competency questions, and related SPARQL queries is required for the hypothesis evaluation process. he input datasets must be queried with all of the competency questions as part of the experiment. The Obtained set of results is made up of the tool's replies to these queries. Inspection and discussion of the query results will form the basis for analysis, as well as comparison to the current ethics literature, with a view on analyzing how this type of project could affect people.

5.4 Results of Competency Questions

The following section details the results which are returned by the User Interface for each competency question.

5.4.1 Query 1:

List Vaccination rates in Liverpool, in order of Date

Local Authority Name	Date	Vaccination Percentage
Liverpool	2020-12-08	0.1
Liverpool	2020-12-12	0.6
Liverpool	...	...
Liverpool	2021-08-12	63.8

Table 5.1: Results for Query 1

There were a total of 248 rows in this response, covering 248 recorded days in the dataset for the Liverpool local authority. The order is oldest date first and most recent date last. This question queries the vaccination dataset. This query gives a foundation to be built on by the other queries. A sample of the response is given in table 5.1.

5.4.2 Query 2:

Get Ethnicity by Local Authority

Local Authority Name	Total	White British	All other White	Mixed or Multiple Ethnic Groups	Asian Asian British	Black/ African/ Black British	Other Ethnic Group
Bury	188	160	6	3	14	3	2
...	...	...	...	...	...	...	...
Wigan	323	306	7	3	4	1	2
Enfield	331	132	70	13	37	60	19

Table 5.2: Results for Query 2

There were 121 rows and 8 columns in this response. The figures for each Ethnicity denote the population in hundreds of thousands for each Ethnicity in each given local authority area. This question queries the ethnicities dataset. This query gives a foundation to be built on by the other queries. A sample of the response is given in table 5.2.

5.4.3 Query 3:

Get Average House price for each Local Authority for each year

Local Authority Name	2019 Average Price	2020 Average Price	2021 Average Price
Barnet	510625.0	544375.0	586250.0
Liverpool	126457.25	132006.25	148625.0
...	...	...	...
Wolverhampton	150937.5	159750.0	170000.0

Table 5.3: Results for Query 3

There were 121 rows and 4 columns in this response. The prices for each year were added up and averaged for each set of prices for each year in each local authority. This question queries the house prices dataset, and some basic manipulation is performed on the data. A sample of the response is given in table 5.3.

5.4.4 Query 4:

List Vaccination rates in Liverpool by date, in reverse order

Local Authority Name	Date	Vaccination Percentage
Liverpool	2021-08-12	63.8
Liverpool	2021-08-11	63.7
...	...	...
Liverpool	2020-12-08	0.1

Table 5.4: Results for Query 4

There were a total of 248 rows in this response, covering 248 recorded days in the dataset for the Liverpool local authority. The order is oldest date first and most recent date last. This competency question queries the vaccination dataset. Some basic manipulation is performed on the data. A sample of the response is given in table 5.4.

5.4.5 Query 5:

Averages of median housing price , Ethnicities, & current Vaccination rate in Liverpool

Local Authority Name	Date	Vacc. Percent	Total	White British	All other White	Mixed or Multiple Eth. Groups	Asian Asian British	Black/ African/ Black British	Other /Ethnic Group	2019 Ave	2020 Ave	2021 Ave
Liverpool	2021-08-12	63.8 487	403	20	9	27	14	14	126457	126457	148625	

Table 5.5: Results for Query 5

There were a total of 4 rows & 13 columns in this response. This competency question queries all three datasets. The prices for each year were added up and averaged for each set of prices for each year in each local authority. The figures for each Ethnicity denote the population in hundreds of thousands for each Ethnicity in each given local authority area. The full response is given in table 5.5

5.4.6 Query 6:

Averages of median housing price, Ethnicities, & current Vaccination rate in Bury

Local Authority Name	Date	Vacc. Percent	Total	White British	All other White	Mixed or Multiple Eth. Groups	Asian Asian British	Black/ African Black British	Other /Ethnic Group	2019 Ave	2020 Ave	2021 Ave
Bury	2021-08-01	73.9	188	160	6	3	14	3	2	172125	179368	197500

Table 5.6: Results for Query 6

This response included a total of four rows and thirteen columns. All three datasets are queried in this competency question. For each set of prices for each year in each local authority, the prices were summed up and averaged. The figures for each ethnicity represent the population of that ethnicity in hundreds of thousands in each local authority region. The full response is given in table 5.6

5.4.7 Query 7:

Averages of median housing price, Ethnicities, & current Vaccination rate in Barking & Dagenham

Local Authority Name	Date	Vacc. Percent	Total	White British	All other White	Mixed or Multiple Eth. Groups	Asian Asian British	Black/ African Black British	Other /Ethnic Group	2019 Ave	2020 Ave	2021 Ave
Barking & Dagenham	2021-08-14	56.9	209	76	26	7	38	51	11	313538	313538	313538

Table 5.7: Results for Query 7

There were four rows and thirteen columns in this response. In this competency question, all three datasets are queried. Prices were totalled and averaged for each set of prices for each year in each local authority. Each ethnicity's data represent the ethnicity's population in hundreds of thousands in each local authority region. The full response is given in table 5.7

5.4.8 Query 8:

Query 8: What is the average Vaccination rate across all local authorities ?

Date	Average Vaccination Rate
2021-12-30	71.474

Table 5.8: Results for Query 8

There were 2 rows and 2 columns in this response. The different local authorities were filtered to have their most recent data, and then all their vaccination rates were averaged. This query serves to put the other vaccination rates quoted previously into some context.

5.4.9 Query 9:

Query 9: What were the ethnicities in the areas with the highest and lowest median house prices ?

Local Authority Name	2021 Average Price	Total	White British	All other White	Mixed or Multiple Eth. Groups	Asian Asian British	Black/ African/ Black British	Other/Ethnic Group
Westminster	882500.0	250	86	60	9	31	19	45
Blackpool	120000.0	140	133	4	1	1	0	1

Table 5.9: Results for Query 9

There were three rows and nine columns in this response. This query links together the Ethnicities dataset and the house prices dataset. Each ethnicity's data represent the ethnicity's population in hundreds of thousands in each local authority region.

5.4.10 Query 10:

Query 10: Get Ethnicity by location as percentages

Local Authority Name	White British	All other White	Mixed or Multiple Ethnic Groups	Asian Asian British	Black/ African/ Black British	Other Ethnic Group
Bury	85.1	3.1	1.5	7.4	1.5	1.0
Milton Keynes	72.5	6.3	2.6	10.5	6.7	1.1
Westminster	34.4	24	3.5	12.4	7.6	18

Table 5.10: Results for Query 10

There were one hundred and twenty one rows and seven columns in this response. This query serves to show some data manipulation of the Ethnicities dataset.

5.5 Observations & Discussion

There are many different ethical aspects to interpreting this data. It is important to acknowledge there are both potential positive, negative and unknown effects of the release of, and manipulation of public datasets. These effects must be more carefully considered by governments releasing more and more datasets everyday.

There are many potentially positive effects of the release of this data, for example Highlighting social inequalities, for example in Query 9 (Table 5.9) we can highlight the differences in ethnicities in rich vs poor areas. Exploration of public datasets can help governments realize that a certain area or population needs more governmental help. Public infrastructure planning. As discussed previously, the release of public data can improve the transparency and accountability of governments, open data holds governments accountable to the results they produce. Residents have the ability to see exactly what their government has achieved, and how much more needs to be done. Failure to achieve specific results or accomplish a specific milestone or target will be reported and scrutinized. Achieving or exceeding goals, on the other hand, can help to build a stronger and more trusted relationship with local citizens. Open government data contributes to citizen confidence and legitimacy. Residents can have confidence that their local government is working hard to keep commitments and make decisions that are in the best interests of the community if they have access to open data. Open governmental data also fosters growth and innovation in academic, public-sector, & industry-based research communities, encourages public education and community engagement, and stores & protects data allowing citizens to track patterns and changes through time.

However despite the many potentially positive effects public data may have, it is also very important to be aware of the many potentially negative effects. One unintended consequence of the release of this data could be increasing Health insurance premiums. For example if a health insurance company can view that there is a particularly low vaccination rate against Covid-19 in a certain area, it is likely that the company will increase the health insurance premiums of the people who live in that area, due to the perceived increased risk of getting more seriously ill if you live in an area with lower vaccination. This is a potentially negative effect on a government's citizens that would not have occurred if the government had kept it's citizen's data private. However it is important to note that there are many determining factors in public health, as illustrated in figure 5.1 below.

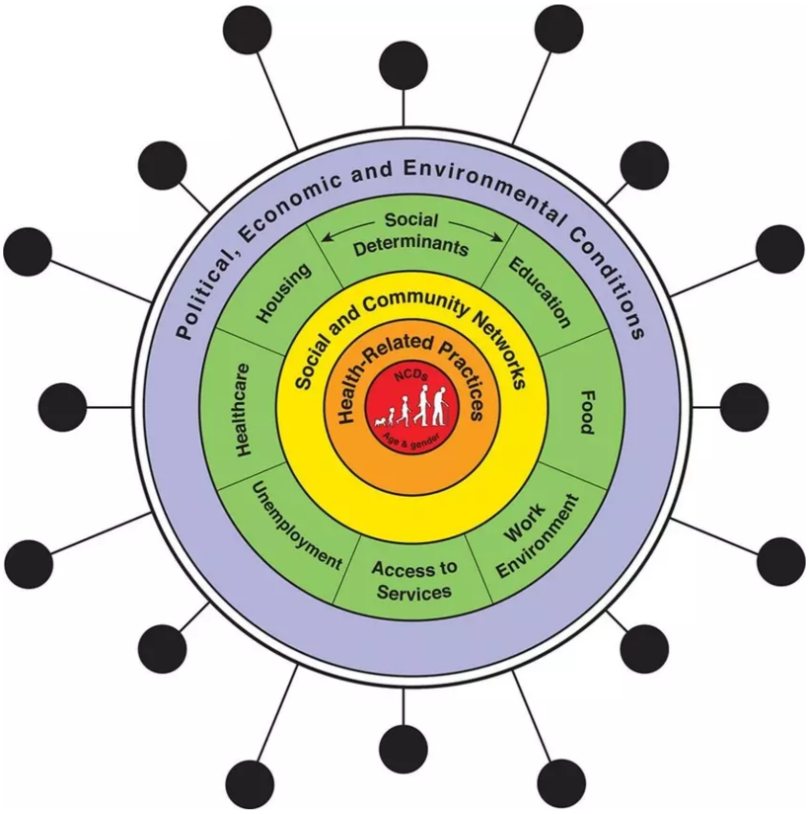


Figure 5.1: The social determinants of health (2)

Exposing public data for commercial re-use has the potential to just serve the interests of capital, further empowering those who are already powerful while disenfranchising others, which relates back to the previously discussed point that public data can serve to empower the empowered. If this is in fact the case, then open data has failed to make society more democratic and open. Public sector data also contains a high degree of social privilege and social values. As a result, data sets contain value systems, which affect analysis and interpretation and serve to perpetuate existing injustices and strengthen dominant interests. (32)

The digitization of land records in Karnataka, India, is a well-known example of "empowering the already empowered," where an open data project promoted as a "pro-poor" initiative actively disenfranchised the poor by allowing those with financial resources and skills to access previously restricted data and reclaim their lands (6). Rather than benefiting all citizens, open data permitted a shift in land rights and a transfer of wealth from the poor to the wealthy in this example. To put it another way, Open data does not imply data democratization. Indeed, open data can function as a tool of disciplinary power

Finally, there is the possibility of a more biased interpretation or manipulation of the results, as well as interpreting results to perpetuate and worsen negative stereotypes in society. This potential for misuse and misinterpretation grows as public data becomes more accessible. Clearly these public datasets are capable of having an impact on everyday people's lives. Therefore, now more than ever, there is a need for greater public ethics awareness.

5.6 Summary

The chapter covers the motivation, goal and hypothesis of the experiment that has been performed by using the case study of linking datasets to explore public data ethics. The case study has been performed using 3 different public datasets to form the use case for further analysis. Result and Analysis along with observations for the experiment and research in general have been discussed.

6 Conclusion

This chapter outlines the degree to which the research objectives have been met. The contributions performed have been stated. A discussion about future work has been made. Eventually, the final remarks have been mentioned.

6.1 Research Objectives

Following research, objectives are derived from the research question:

RO1: *Survey of state of the art on Open data, data ethics and linked data.*

The objective of reviewing the existing literature and publicly available tools to Link public datasets was carried out. And also reviewing the current literature on Open data & data ethics was successful and a well rounded background was established.

RO2: *Obtaining datasets which could potentially be linked together in what could be considered an ethically questionable manner*

Several datasets were found which had a suitable ethical component for this project and contained sufficient similarities. These datasets were then used to form a Use case to investigate the ethics of Linked Open data.

RO3: *Developing an approach to link the datasets together.* An approach to link the datasets was developed using a designed ontology, an R2RML mapping, GraphDB as a triplestore and finally a python based User interface to query the linked database.

RO4: *Developing a tool which would display the some "Competency questions" online to query the linked database.*

The competency questions were displayed in a interactive User interface, and they displayed some questions which highlighted potentially ethically questionable areas of the data linkage.

RO5: *Evaluate the results and potential ethical and societal impacts.*

The societal and ethical impacts have been discussed in the results section.

6.2 Contributions

1. A concrete Use Case was developed using Linked Open data to discuss public data & linked data ethics.

2. A general discussion on public data ethics and the potential dangers that the free release of these datasets can have.

6.3 Technical Difficulties

A deeper knowledge in ethics would have been hugely beneficial to this project. If we were to redo the research we would consider including a social scientist. The findings were also limited by the limited number of datasets, as linking more datasets could have strengthened the use case and given other aspects of data ethics to discuss.

6.4 Future Work

The use case present in this report serves to illustrate the dangers of public data, therefore, more thought needs to be put into how Linked Data could affect people and their lives. Given the Implications of this research, we must be more aware and put more thought into how the release of public date could affect people and their lives.

In terms of future work, more study could be done to investigate how to assess & measure the consequences and repercussions of releasing public statistics. Something that has been discussed as a serious challenge in the current literature thus far. At the very least more thought into the potential repercussions of the release of these datasets must be put in by governments. A recommendation for future work would be implementing some "Data Ethics warning" or popup which is featured one government websites whenever someone downloads a dataset. This warning could contain a warning on the potential harms which could come from the misuse of public data, as well as some links for further reading .

6.5 Final Remarks

This dissertation explores public data, its uses and potential effects on wider society. When dealing with public data, many governments are quick to point out all of the potentially beneficial effects, and less inclined to explore the potential dangers of public data. As discussed public data can be a force for good in building trust, transparency & accountability in governments. It can have many other positive impacts on society, therefore this paper does not seek to discourage the release of public data. However, society must be aware of the negative aspects and dangers of this data, as it can be a very powerful tool. As previously established, there are numerous negative consequences of public data that we must be cautious of. Public data can be used to Empower those already in power, which can perpetuate existing injustices and strengthen dominant societal and cultural values which may be ethically questionable. It is clear that as society progresses to be increasingly data driven, we need to be aware of data ethics, and its role in our society.

Bibliography

[1] 2022. URL `https://www.ontotext.com/knowledge-hub/fundamentals`.

[2] Clare Bambra, Ryan Riordan, John Ford, and Fiona Matthews. The covid-19 pandemic and health inequalities. *J Epidemiol Community Health*, 74(11):964–968, 2020.

[3] Department of Public Expenditure and Government Reform Unit Reform. Open data strategy 2017 – 2022, 2022. URL `https://data.gov.ie/uploads/page_images/2018-03-07-114306.063816Final-Strategy-online-version1.pdf`.

[4] 2022. URL `https://www.gov.uk/government/publications/uk-national-data-strategy/national-data-strategy`.

[5] Michael Chui, Diana Farrell, and Kate Jackson. How government can promote open data. *McKinsey Company*, 2014.

[6] Michael B Gurstein. Open data: Empowering the empowered or effective data use for everyone? *First Monday*, 2011.

[7] Anneke Zuiderwijk and Marijn Janssen. The negative effects of open government data-investigating the dark side of open data. In *Proceedings of the 15th Annual International Conference on Digital Government Research*, pages 147–152, 2014.

[8] Tim Berners-Lee, James Hendler, and Ora Lassila. The semantic web. *Scientific american*, 284(5):34–43, 2001.

[9] Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data: The story so far. In *Semantic services, interoperability and web applications: emerging concepts*, pages 205–227. IGI global, 2011.

[10] Bernhard Haslhofer and Antoine Isaac. data. europeana. eu: The europeana linked open data pilot. In *International Conference on Dublin Core and Metadata Applications*, pages 94–104, 2011.

[11] Tim Berners-Lee. Linked data - design issues, 2006. URL `https://www.w3.org/DesignIssues/LinkedData.html`.

[12] 2022. URL https://www.ontotext.com/knowledgehub/fundamentals/linked-data-linked-open-data/.

[13] Kingsley Okoye. Linked open data: State-of-the-art mechanisms and conceptual framework. In *Linked Open Data-Applications, Trends and Future Developments*. IntechOpen, 2020.

[14] Alasdair MacIntyre. *A short history of ethics: a history of moral philosophy from the Homeric age to the 20th century*. Routledge, 2003.

[15] Merel Noorman. Computing and moral responsibility. 2012.

[16] Gottfried Wilhelm Leibniz. *Theodicy: Essays on the Goodness of God, the Freedom of Man and the Origin of Evil*. Wipf and Stock Publishers, 2000.

[17] Luciano Floridi. Network ethics: Information and business ethics in a networked society. *Journal of business ethics*, 90(4):649–659, 2009.

[18] Shannon Vallor. Social networking and ethics. 2012.

[19] Toni Erskine. *Can institutions have responsibilities?: collective moral agency and international relations*. Springer, 2003.

[20] Terrell Bynum. Computer and information ethics. 2001.

[21] J Hurwitz, Alan Nugent, Fern Halper, and Marcia Kaufman. Big data. *New York*, 2013.

[22] Andrej Zwitter. Big data ethics. *Big Data & Society*, 1(2):2053951714559253, 2014.

[23] K Cukier. Kenneth cukier (data editor, the economist) speaks about big data, 2013.

[24] Luciano Floridi. Soft ethics, the governance of the digital and the general data protection regulation. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2133):20180081, 2018.

[25] Luciano Floridi. Technoscience and ethics foresight. *Philosophy & Technology*, 27(4): 499–501, 2014.

[26] Mohammad Alamgir Hossain, Yogesh K Dwivedi, and Nripendra P Rana. State-of-the-art in open data research: Insights from existing literature and a research agenda. *JOURNAL OF ORGANIZATIONAL COMPUTING AND ELECTRONIC COMMERCE*, 26(1-2):14–40, 2016.

[27] 2022. URL https://www.w3.org/RDF/.

[28] 2022. URL https://www.w3.org/2001/sw/wiki/OpenLink_Virtuoso.

[29] Lambrini Seremeti and Achilles Kameas. Tools for ontology engineering and management. In *Theory and Applications of Ontology: Computer Applications*, pages 131–154. Springer, 2010.

[30] 2022. URL `https://www.w3.org/wiki/LargeTripleStores`.

[31] 2022. URL `https://www.local.gov.uk/about/what-local-government`.

[32] Prof. Rob Kitchin. Four critiques of open data initiatives, 2013. URL `https://blogs.lse.ac.uk/impactofsocialsciences/2013/11/27/four-critiques-of-open-data-initiatives/`.

A0

A Appendix

A.1 SPARQL Queries

Query 1: List Vaccination rate in Liverpool , in order of Vaccination rate

```
PREFIX onto:<http://www.ontotext.com/>
PREFIX jk:
<http://www.semanticweb.org/thompsha/ontologies/2022/2/disertationOntology>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT DISTINCT ?vaxNames ?date ?cumVaccinationPercent
WHERE {
    ?vax a jk:vaxNums.
    ?vax rdfs:label ?vaxNames.
    ?vax jk:date ?date.
    ?vax jk:cumVaccinationSecondDoseUptakeByVaccinationDatePercentage
?cumVaccinationPercent .
FILTER (?vaxNames = 'Liverpool')
}
ORDER BY ?vaxNames LIMIT 500")
```

Query 2: Get Ethnicity by location PREFIX onto:<http://www.ontotext.com/>

```
PREFIX jk:
<http://www.semanticweb.org/thompsha/ontologies/2022/2/disertationOntology>

PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT DISTINCT ?vaxNames ?Total ?x ?x2 ?x3 ?x4 ?x5 ?x6

WHERE {

    ?localAuth a jk:localAuthEthnicity . ?localAuth rdfs:label ?vaxNames .

    ?localAuth jk:Total ?Total .

    ?localAuth jk:White British ?x.

    ?localAuth jk:All Other White ?x2 .

    ?localAuth jk:Mixed OR Multiple ethnic groups ?x3 .

    ?localAuth jk:Asian OR Asian British ?x4 .

    ?localAuth jk:Black OR African OR Caribbean OR Black British ?x5 .

    ?localAuth jk:Other ethnic group ?x6 . }
```

Query 3: Get Average House price for each Local Authority for each year

```
PREFIX onto:<http://www.ontotext.com/>

PREFIX jk:
<http://www.semanticweb.org/thompsha/ontologies/2022/2/disertationOntology>

PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT ?locName (((?Mar 2019+?Jun 2019+?Sep 2019+?Dec 2019)/4) AS ?avg2019)
(((?Mar 2020+?Jun 2020+?Sep 2020+?Dec 2020)/4) AS ?avg2020) (((?Mar 2021 + ?Jun
2021)/2) AS ?avg2021)
WHERE {

    ?med a jk:medianHousePrice.

    ?med rdfs:label ?locName.

    ?med jk:Year ending Mar 2019 ?Mar 2019.

    ?med jk:Year ending Jun 2019 ?Jun 2019.

    ?med jk:Year ending Sep 2019 ?Sep 2019.

    ?med jk:Year ending Dec 2019 ?Dec 2019.

    ?med jk:Year ending Mar 2020 ?Mar 2020.

    ?med jk:Year ending Jun 2020 ?Jun 2020.

    ?med jk:Year ending Sep 2020 ?Sep 2020.

    ?med jk:Year ending Dec 2020 ?Dec 2020.

    ?med jk:Year ending Mar 2021 ?Mar 2021.

    ?med jk:Year ending Jun 2021 ?Jun 2021. }

order by ?locName")
```

Query 4: Rank Vaccination rate in Liverpool in reverse order

```
PREFIX onto:<http://www.ontotext.com/>
PREFIX jk:
<http://www.semanticweb.org/thompsha/ontologies/2022/2/disertationOntology>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT DISTINCT ?vaxNames ?date ?cumVaccinationPercent
WHERE {
    ?vax a jk:vaxNums.
    ?vax rdfs:label ?vaxNames.
    ?vax jk:date ?date.
    ?vax jk:cumVaccinationSecondDoseUptakeByVaccinationDatePercentage
?cumVaccinationPercent .
FILTER (?vaxNames = 'Liverpool')
}
ORDER BY ?date LIMIT 500")
```

Query 5: Averages of median housing price , Ethnicities, & current Vaccination rate in Liverpool

```
PREFIX onto:<http://www.ontotext.com/>

PREFIX jk:
<http://www.semanticweb.org/thompsha/ontologies/2022/2/disertationOntology>

PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT DISTINCT ?vaxNames ?Total ?x ?x2 ?x3 ?x4 ?x5 ?x6

WHERE {

    ?localAuth a jk:localAuthEthnicity .

    ?localAuth rdfs:label ?vaxNames .

    ?localAuth jk:Total ?Total .

    ?localAuth jk:White British ?x.

    ?localAuth jk:All Other White ?x2 .

    ?localAuth jk:Mixed OR Multiple ethnic groups ?x3 .

    ?localAuth jk:Asian OR Asian British ?x4 .

    ?localAuth jk:Black OR African OR Caribbean OR Black British ?x5 .

    ?localAuth jk:Other ethnic group ?x6 .

FILTER (?vaxNames = 'Liverpool')}
```

Query 6: Averages of median housing price, Ethnicities, & current Vaccination rate in Bury

```
PREFIX onto:<http://www.ontotext.com/>

PREFIX jk:
<http://www.semanticweb.org/thompsha/ontologies/2022/2/disertationOntology>

PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT DISTINCT ?vaxNames ?Total ?x ?x2 ?x3 ?x4 ?x5 ?x6

WHERE {

    ?localAuth a jk:localAuthEthnicity .

    ?localAuth rdfs:label ?vaxNames .

    ?localAuth jk:Total ?Total .

    ?localAuth jk:White British ?x.

    ?localAuth jk:All Other White ?x2 .

    ?localAuth jk:Mixed OR Multiple ethnic groups ?x3 .

    ?localAuth jk:Asian OR Asian British ?x4 .

    ?localAuth jk:Black OR African OR Caribbean OR Black British ?x5 .

    ?localAuth jk:Other ethnic group ?x6 .

FILTER (?vaxNames = 'Bury')}
```

Query 7: Averages of median housing price, Ethnicities, & current Vaccination rate in Barking & Dagenham

```
PREFIX onto:<http://www.ontotext.com/>
PREFIX jk:
<http://www.semanticweb.org/thompsha/ontologies/2022/2/disertationOntology>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT DISTINCT ?vaxNames ?Total ?x ?x2 ?x3 ?x4 ?x5 ?x6
WHERE {
    ?localAuth a jk:localAuthEthnicity .
    ?localAuth rdfs:label ?vaxNames .
    ?localAuth jk:Total ?Total .
    ?localAuth jk:White British ?x.
    ?localAuth jk:All Other White ?x2 .
    ?localAuth jk:Mixed OR Multiple ethnic groups ?x3 .
    ?localAuth jk:Asian OR Asian British ?x4 .
    ?localAuth jk:Black OR African OR Caribbean OR Black British ?x5 .
    ?localAuth jk:Other ethnic group ?x6 .
FILTER (?vaxNames = 'Barking & Dagenham')}
```

Query 8: Average Vaccination rate across all local authorities

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#> PREFIX
onto:<http://www.ontotext.com/> PREFIX jk:
<http://www.semanticweb.org/thompsha/ontologies/2022/2/disertationOntology>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT DISTINCT ?date (AVG(?cumVaccinationPercent) AS ?AVGcumVaccination)

WHERE {

    ?vax a jk:vaxNums .

    ?vax jk:date ?date.

    ?vax jk:cumVaccinationSecondDoseUptakeByVaccinationDatePercentage
?cumVaccinationPercent .

FILTER (?date= '2021-12-30') } GROUP BY ?date ?cumVaccinationPercent
```

Query 9: What were the ethnicities in the areas with the highest and lowest median house prices ?

```
PREFIX onto:<http://www.ontotext.com/>

PREFIX jk:
<http://www.semanticweb.org/thompsha/ontologies/2022/2/disertationOntology>

PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT DISTINCT ?vaxNames ?Total ?x ?x2 ?x3 ?x4 ?x5 ?x6

WHERE {      ?localAuth a jk:localAuthEthnicity .

      ?localAuth rdfs:label ?vaxNames .

      ?localAuth jk:Total ?Total .      ?localAuth jk:White British ?x.

      ?localAuth jk:All Other White ?x2 .      ?localAuth jk:Mixed OR Multiple ethnic
groups ?x3 .

      ?localAuth jk:Asian OR Asian British ?x4 .

      ?localAuth jk:Black OR African OR Caribbean OR Black British ?x5 .

      ?localAuth jk:Other ethnic group ?x6 .

FILTER (?vaxNames = 'Blackpool')}

UNION

SELECT DISTINCT ?vaxNames ?Total ?x ?x2 ?x3 ?x4 ?x5 ?x6

WHERE {      ?localAuth a jk:localAuthEthnicity .

      ?localAuth rdfs:label ?vaxNames .

      ?localAuth jk:Total ?Total .      ?localAuth jk:White British ?x.

      ?localAuth jk:All Other White ?x2 .      ?localAuth jk:Mixed OR Multiple ethnic
groups ?x3 .

      ?localAuth jk:Asian OR Asian British ?x4 .

      ?localAuth jk:Black OR African OR Caribbean OR Black British ?x5 .

      ?localAuth jk:Other ethnic group ?x6 .

FILTER (?vaxNames = 'Westminster')}
```

Query 10: Get Ethnicity by location as percentages

```
PREFIX onto:<http://www.ontotext.com/> prefix jk:
<http://www.semanticweb.org/thompsha/ontologies/2022/2/disertationOntology>

PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT DISTINCT ?vaxNames ?Total ?x ?x2 ?x3 ?x4 ?x5 ?x6 WHERE {

    ?localAuth a jk:localAuthEthnicity .

    ?localAuth rdfs:label ?vaxNames .

    ?localAuth jk:Total ?Total .

    ?localAuth jk:White British ?x.

    ?localAuth jk:All Other White ?x2 .

    ?localAuth jk:Mixed OR Multiple ethnic groups ?x3 .

    ?localAuth jk:Asian OR Asian British ?x4 .

    ?localAuth jk:Black OR African OR Caribbean OR Black British ?x5 .

    ?localAuth jk:Other ethnic group ?x6 .

}
```