Constructing a Knowledge Graph for Open Statistical Data

Enayat Rajabi, Rishi Midha *, Devanshika Ghosh
Cape Breton University, Sydney, NS, Canada
E-mail: {enayat_rajabi, cbu19ffj, cbu19cgx}@cbu.ca

Abstract. Open Government Data has been published widely by different governments to be used by public and data consumers. The majority of published datasets are statistical. Transforming Open Government Data into Knowledge graphs bridge the semantic gap and give machines the power to logically infer and reason. Through this paper, a knowledge graph is proposed for Open Statistical Data. An RDF-based knowledge graph with a rule-based ontology are presented on this paper. A case study on Nova Scotia Open Data (a provincial Open Data portal in Canada) is also presented. The proposed knowledge graph can be used on any statistical Open Data and can bring all provincial Open Government Data under a single umbrella. The knowledge graph was tested and underwent a quality check process. The study shows that the integration of statistical data from multiple sources using ontologies and interlinking features of Semantic Web potentially enables the performance of advanced data analytics and leads to the production of valuable data sources and it generates a dense knowledge graph with cross-dimensional information and data. The ontology designed to develop the graph adheres to best practices and standards thereby allowing for expansion, modification and flexible re-use.

Keywords: Open Statistical Data, Knowledge Graph, Semantic Web, Ontology, Linked Data

1. Introduction

Publishing open statistical data is getting increased attention on the Web [1] and the various levels of government. The so-called open government data movement has led to the launch of open data portals with providing a single point of access for a province or country. Open government data increases government transparency, accountability, contributes to economic growth and improves administrative processes [2]. The open government data is published to the hope that it can be used by different organizations, data consumers in public and private sectors. For instance, the tracking of a global pandemic has been made possible by the use of open data published by regional authorities. This has not just aided government and administrative institutions but has also helped investigate and to a certain degree control the spread of the outbreak. The use of open data can be seen in the business sector as part of market studies for identifying demographics, consumer information and product development. As this data is freely available and accessible, it can be used by small and large establishments alike, thereby helping bridge the gap and helping diminish disparity. In addition, the publishing of open data further advances the concept of knowledge creation through reuse and collaboration. The majority of open government data, however, consists of statistical data published by governments around the world [3]. A variety of open statistical data are published such as census data, demographics, public health data (e.g., number of disease cases) or business data (number of unemployment), which can be used in public services and provides social value to citizens [4]. In itself, the data can be restrictive and not powerful enough to draw meaningful inferences from. The datasets act as isolated pools of information that cannot be queried or linked. These sources are scattered and the information can only be accessed from individual datasets through specific search. The lack of meaning behind the data makes it impossible to form a network and link this open statistical data to infer, create and query knowl-

*Corresponding author. E-mail: cbu19ffj@cbu.ca.

edge. By transforming them into linked datasets by generating a knowledge graph, information is inferred, created and queried.

Open statistical datasets can be linked to each other to bring about enrichment of knowledge to the dataset. As a result, data becomes complete and more meaningful. For instance, by combining datasets containing government funding in various departments such as agriculture, housing etc one can keep track of all government funding. Transforming open statistical data to knowledge graphs can also enable performing analytics on top of disparate and previously isolated datasets in an open government portal [4]. Interconnectivity between isolated open datasets gives a machine a lot of information to work with and thereby strengthen its ability to deduce relations and infer meaning. For the data across fields as diverse as finance, medicine and arts to be linked there has to be common grounds and some degree of homogeneity and uniformity. It is possible to link various open datasets from a number of resources and gain insights or knowledge which was not possible with a single source. Open statistical data are often published by different organizations pertaining to a particular subject. To make different data compatible with each other, some standards should be utilized to regularize this process. The data are usually defined by ontologies to unify this approach and to ultimately use machines to draw new inferences that enrich the information available to users. The ontology in a knowledge graph focuses on linking arbitrary entities or concepts from different domains using classes and relationships. The design of knowledge graph also ensures minimal usage of literals or strings when modelling statistical data which seems bizarre but is possible when extended to dimensional modelling of said data. From a broader perspective, any graph-based representation based on ontologies and RDF that can create knowledge, could be considered a knowledge graph [5]. This study focuses on forming complex networks of diverse domains using links and semantic relations to generate a dense knowledge graph enriched by the use of semantic rules and axioms. This adds a richer meaning and new layer to the existing semantic relations and makes possible the extraction of even more information from data. The research question of this study is: "How can one infer new knowledge from open statistical data using semantic web technologies?" To answer the above mentioned question, ontologies, RDF multidimensional models, and deductive reasoning rules were utilized to generate a knowledge graph for open statistical data. For instance, this study used several disease datasets in one open data portal and linked them to the disease ontology. By doing so one can not just query the patients of a certain disease but query the cases of a general health category (e.g., viral infection). This amalgamated different datasets only because of linking of datasets and was possible as the study was able to present this linking in a machine readable form. Hence, the system was in itself able to infer knowledge and enhance not only its knowledge base but also user experience.

The use of existing ontologies, vocabularies, and standards such as rdf cube are leveraged to create a graph with semantic relations and properties of typed entities. The instances of statistical data are linked semantically on a schema-level basis which is a criterion for creation of knowledge graph. This relationship is obtained by following cube vocabulary and well-known ontologies for expressing other metadata, forming dimensional links and using rules to add another layer to the relationship between entities. The use of objects instead of literals also promotes links between triples that adds another complexity to the graph thereby enriching the entire network. After creating the knowledge graph, a quality and refinement process was performed using a specific quality metric to measure the accuracy and precision of the created knowledge graph and its entities and classes based on refinement standards [6] and [7].

The structure of this paper is as follows: Section 2 discusses the related studies along with the multidimensional model that we used for generating the knowledge graph. Section 3 outlines the architecture, model, and ontologies we used in this paper to create the knowledge graph. Section 4 explains the case study which is open government data, followed by conclusion in Section 5.

2. Background

Languages, such as RDF and OWL, are used in an ontology to represent the semantics of an entity as a set of things or concepts rather than strings of words [8]. They provide rich constructs to represent information that is machine understandable and facilitate semantic integration and sharing of information from heterogeneous sources [9]. Ontologies, as human- and computer interpretable representations of the types of entities, have gained a lot of popularity and recognition in the semantic web because of their extensive use in Internet-based applications. They are of-

ten considered a fine source of semantics and interoperability in all artificially smart systems [10]. Most of data on the Web are stored in relational databases and are accessible for humans through Web browsers. Extraction of the database schema and representing it as ontology ensure semantic access to database data. There are a grand variety of methods for exposing relational database to semantic web, differing from each other in used models (annotation or translation), languages and additional database manipulation techniques. Some extract the schema from the database and convert it to semantic web format; others use annotations, or wrappers. A knowledge graph is built on top of a inter-connected network of entities of these ontologies. The entities comprising a knowledge graph have semantic relations and dimensional links making traversal through the dense graph easy. This gives rise to the inference and creation of knowledge and facts which relational data alone cannot make available. This knowledge is readily available to be queried and re-used. To construct a knowledge graph for an open statistical data, a multidimensional structure should be defined. Statistical data consists of measures (e.g., number of cases) and dimensions describing the measures (e.g., regions). There is a growing need for the statistical data to be published in a way that it can be linked to related datasets and combined with associated information (metadata included). The RDF Data Cube vocabulary is a W3C recommendation and an efficient solution to this need as it enables representation of the statistical data in standard RDF format and publishes the data conforming to the principles of linked data which allows entities used in the datasets to be linked with other related or linked datasets hence improving discoverability, reuse, and, sharing [11]. This vocabulary is an ontology used to describe multi-dimensional datasets, while supporting and building upon extension RDF vocabularies such as Statistical Data and Metadata eXchange XML format (SDMX), SKOS, SCOV1O, Dublin Core, FOAF, etc which augments additional context to the statistical data. The RDF Cube vocabulary has been widely used in different studies [12], [13] ,[14] and accepted by the Semantic Web community, as it allows publishers to integrate and slice across their datasets. [15] aptly describes the structure of the dataset using the vocabulary in their research paper. According to it, structure of the dataset is defined by a Data Structure Definitions (DSDs) where observations are identified by their dimensions, capture one or more observed values via measures, and observed values can be annotated with attributes.

Literature cites many examples where researchers and organizations alike have implemented the RDF Data Cube vocabulary. [14] describes the process of improving and enriching the quality of published data by means of multi-dimensional data, applying linked open data assessment process and using external repositories as a knowledge base. This case study was applied to Barcelona's official open data platform to great effect. In another example, [16] in their paper described how the Czech Social Security Administration (CSSA) published their official pension statistics as linked open data (LOD). These LOD datasets were modelled using the Simple Knowledge Organization System (SKOS) vocabulary and the RDF Data Cube Vocabulary (DCV). The survey paper [17], highlights the importance of ontology to coding knowledge for treatment and development of the Dementia-Related Agitation Non-Pharmacological Treatment Ontology (DRANPTO). As a knowledge base, it plays a vital role to facilitate the development of intelligent systems for managing agitation in dementia. Highly domain specific in its implementation, the ontology applies the concepts of resource re-use and adaptability to create a knowledge base and aims to elicit knowledge on dementia.The use of open statistical data in the medical industry, in the form of health reports, medical reports and other such data, have been used in the study. Another ontology in the health IT interventions domain, developed and published in the study [18], builds on existing pre-existing health and medical ontologies. The study outlines an inductive-deductive approach to develop a glossary and define concepts such as classes and instances, and finally publish the ontology as linked open data. In contrast, this study sweeps a larger domain base and aims to integrate multi-faceted statistical data to build a knowledge base and subsequently answer a variety of questions. Research in the domain of business led to development of the euBusinessGraph Ontology which collects and harmonizes business company information. The study of this paper is similar to this research by following W3C standards, promoting re-use and adaptability, mapping data and datasets and using quality and refinement techniques for analysis. In the field of education, the studies [19] and [20] outline the modelling of an ontology and knowledge graph respectively. The former comprises of curriculum and syllabus and uses open statistical data from the education industry to populate said ontology. The framework to design the ontol-

ogy and define the core concepts and promote linked data concepts is in coherence with this study. The latter focuses on modelling knowledge graph for heterogeneous data automatically by identifying subjects and educational relations through association rule mining and neural sequence labelling.

The PubMed knowledge graph [21] is created from the PubMed library. The study outlines the extraction of over 29 million records from the library to generate a graph that links bio-entities, authors, funding, affiliations and articles. Subsequent data validation yielded promising results and the graph is able to create and transfer knowledge, profile authors and organizations and realize meaningful links between bio-entities. The study covers familiar territory in terms of knowledge graph and generation when compared to the work done in this research study. The area of research pertaining to war casualties, as outlined in the study by [22], details modelling of knowledge graph for open statistical data about world war two casualties. The study aims to link this data with the existing WarSampo system. This not only helps provide historical records for public access through publishing of aforementioned records as linked data, but it also aids digital humanities research. The study details extensive depth and granularity as it integrates the approximate 100,000 records available in a relational database at the National archives of the Finnish army. The dataset is available as a SPARQL endpoint. The knowledge graph studies various cases such as studying death records and reassembling soldier biographies and military unit histories to answer complex questions. The study [23] explores the creation of a knowledge graph called DBkWik, a complimentary knowledge graph. The famous knowledge graphs DBpedia and Yago form the basis for further study in the field. Wiki farms like Fandom contain specific information often complimentary to that available in Wikipedia. The study outlines the generation and data validation performed as well as certain challenges and domain independence that resonate with the study outlined in this paper. The literature and research in the field has been promising. The potential of LSOD is yet to be exploited in its entirety. The aim of this study is to move one step closer to harnessing that potential. The study contrasts from other works in the sense that it is not domain specific. The aforementioned studies involved research and work on domain specific data whereas this study has attempted to design a primer that fits statistical data regardless of domain. This greatly benefits in providing homogeneity and compatibility across domains thus creating the possibility of creation of a rich knowledge base and subsequently a dense knowledge graph. This research also does not depart too far from other works in the field. The use of standards and strict adherence to well-established rules and protocols of the semantic web prescribed by W3C ensure compatibility with past works as well.

3. Methodology

The research aims to build a knowledge graph on open statistical data. The systematic approach followed to build a knowledge graph and complete the study can be broken down into methodical steps. These steps were adopted after review of current literature as well as best practices in the field.

3.1. Data Collection

The first step pertains to the collection of raw data. The data can be obtained from a variety of web hosted services. The statistical datasets are usually stored as records with multiple dimensions and measures. The dimensions of data reflect the detail or level of abstraction. They dictate the level of view and help in categorization or segmentation of records. The measures reflect the values being measured or studied. The data, however, lacks the semantics to be linked with other data. It needs to be given unique identifiers and meaning through semantics, logic and linking which forms the basis for subsequent steps.

3.2. Designing Ontology

To generate a knowledge graph, the extracted datasets should be converted to RDF triples. This requires to design an ontology that can accommodate said data. The use of the standards prescribed by W3C was selected as the bedrock for this study. Other ontologies with domain knowledge to express dimensions from selected data were also imported to the knowledge graph.

3.3. Enriching Data

A statistical open dataset comprises a collection of observations made at some points across some logical space. Each dataset, a collection of multiple records, forms an entity. The data is transformed into RDF or semantic triples to express the dimensions and

measures. The dimensions were also adopted from external ontologies. This helps add domain knowledge to data resulting in an increased knowledge base and more inferred information. The inter-linking of datasets through use of common structure was carried out. This enriched the datasets to create a sound knowledge base.

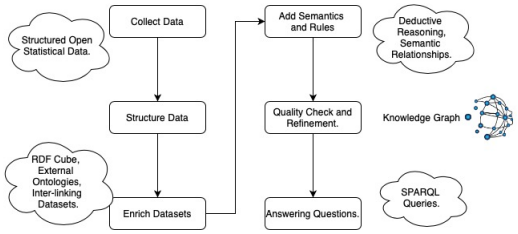

Fig. 1. A Flow Diagram Of The Methodology

3.4. Adding Semantics and Rules

The entities are rigorously inter-linked and engaged in semantic relationships with other entities to form the graph. Linked data serialized as RDF also inherently lacks complex formal semantics that would allow a reasoner to infer the relationship between data items in different datasets [24]. A knowledge graph necessitates inter-linking of entities and use of concepts such as classes, properties and instances to form semantic relationships. This step was carried out to add more meaning to data and form a dense knowledge graph. The use of rules through built-in Protege[1] tabs added another layer of complexity to the graph. This helps add another semantic layer and links the data together. The rules however are undecidable. This means that they may reinforce logic but cannot be solely used for adding semantics to data. The best practice is to re-use and incorporate existing vocabularies and ontologies so that the data can become part of the linked data pool on the web.This was carried out in preceding steps.

3.5. Quality Check and Refinement

A knowledge graph must be able to generate information with a certain degree of integrity. There must be validity and accuracy in the information retrieved. The study tests the validity and correctness of the knowledge graph using techniques elicited in modern well-known research as well as specifically developed software tools. A review of tools and metrics used as industry standards was carried out. The metrics used for quality assessment of a knowledge graph include common measures such as accuracy and precision as well as certain case-specific parameters. The process of refinement ensures that the knowledge graph, when queried, produces authentic information. A review of tools and metrics used as industry standards was carried out to identify the relevant source material for this study.

3.6. Designing Questions and SPARQL Queries

The knowledge graph is only justified if it is able to provide information that cannot be accessed by the web portal. The motivation of the study was to go beyond the capabilities of current web portals through use of semantics to be able to obtain meaningful results. The questions were designed keeping in mind the rules and semantics as well as a certain degree of complexity. Subsequently, to answer these questions, SPARQL Queries were designed to retrieve and manipulate the data stored in a knowledge graph.The queries were designed such that they leverage the full potential of the rules to truly demonstrate the power of the semantic web. Cross-validation of results to expose shortcomings and errors was also carried out to further aid refinement.

4. Case Study: Nova Scotia Open Data Portal

It was observed that leading administrations all around the world were actively involved in making most data available to the public to view and understand. One similar open data portal is the Nova Scotia Open Data portal which was launched in February 2016 as an initiative taken by the government to make government's data publicly accessible. The primary goal of this portal is to provide open access to the government's data to exemplify transparency, and to empower citizens, businesses and researchers to use the province's collective knowledge to innovate ways for social and economic growth. This portal has data which has been collected from multiple sectors and domains and this is why we have constructed and designed our ontologies using this data. The data collections in the portal are periodically refreshed and new data is also added regularly. At the time of writing this research, there are over 500 datasets from various public sectors with the most accessed datasets from public

[1] https://protege.stanford.edu

service information, historical vital statistics and environmental monitoring reports (NovaScotia.ca) in Nova Scotia province in Canada.

To obtain the data, a command line tool was built that fetches data from all datasets of Nova Scotia data portal. This tool has been programmed using Node JS. Socrata Open Data API was used to access the repository using code. The Socrata Open Data API allows you to programmatically access a wealth of open data resources from governments, non-profits, and NGOs around the world [2]. To understand the data further, we performed an exploratory analysis to obtain a general insight about the information contained in the datasets and conduct a more detailed research that shall discover some interesting patterns and similarities between them. These patterns and similarities will help us in creating better ontologies. This data analysis was conducted using Python programming language. There were a total of 518 datasets grouped into Archived and Currently Active categories. At the time of this research, 77.8% were Archived datasets and 22.2% were Currently Active. There were a total of 28 unique categories in NSOD. The top categories were Environment and Energy (58), Health and Wellness (52), Population and Demographics (48), Business and Industry (37) and Education – Primary to Grade 12 (32).

All the datasets were assigned a set of tags. There were a total of 1,154 tags and the highest tag was Community counts (111), followed by Census (42), Canadian ambient air quality and pollution (111), environment (74), and education (27). The majority of datasets were created from April 2016 to June 2016. In 2020, around 29 datasets were created and 197 datasets have been updated. In terms of language, the majority of datasets are in English. Figure X below shows the distribution of the datasets with and without Nova Scotia as the region. Around 79.7% of the datasets have Nova Scotia defined as their region, while 20.3% datasets have missing values in region metadata.

4.1. Knowledge Graph for Open Statistical Data

To generate a knowledge graph for Nova Scotia Open Data Portal, we gathered data and transformed to RDF triples. This enabled inclusion in the ontology. The ontology was then extensively processed to enrich data through internal and external linking as well as dimensional and logical relations. The criteria of a knowledge graph were implemented through rules and inter-linking using semantic relationships. The graph was subjected to a quality and refinement check. This is followed by query retrieval to answer complex questions using SPARQL. The entire process required extensive study of contemporary literature as well as design of elements such as rules and semantic relationships to create a sound knowledge base for the graph.

4.1.1. Ontology design

A Dataset is a collection of statistical data that corresponds to a defined structure. All the datasets in the ontology are all instances of class qb:DataSet. The nomenclature used for datasets is stat:dataset-name. It includes all the metadata specified by DCMI [3] and RDFS to improve readability. Each dataset is a collection of dimensions, measures and attributes. The dimensions act as unique identifiers i.e. with values for each dimension, a record can be identified. The measures are the actual values observed across logical space, time, or any other dimension. The attributes quantifies the measurement and helps in its interpretation. The dimension, measures and attributes of a dataset together comprise its structure and are thus aptly stored in what is called the Data Structure Definition. It may define the structure of one or more datasets, however, each dataset can have only one associated data structure definition. The nomenclature is specified as stat:dsd-name. It is an instance of qb:DataStructureDefintion and linked with qb:DataSet by property qb:structure. The dimensions, measures and attributes are also instances of sub-classes of qb:ComponentProperty, namely qb:DimensionProperty, qb:MeasureProperty and qb:AttributeProperty respectively. They are linked with the Data Structure Definition by properties qb:dimension, qb:measure and qb:attribute respectively. A common occurrence in statistical datasets is presence of code lists which are specified as qb:CodedProperty. The use of existing standards sdmx-code to promote re-usability and homogeneity has been followed.

The dimensions can further be used to form logical slices of the cube structure of datasets. Instances of class qb:Slice and sub-class of qb:ObservationGroup, they are used to group observations by fixing the value of one or more dimensions. Each slice has an associated qb:SliceKey linked to it by qb:sliceStructure. The property qb:slice links the slice to the dataset

[2] https://dev.socrata.com/

[3] https://dublincore.org

Ontology	Usage
RDF Cube	The use of classes, properties and concepts to represent datasets
Dublin Core Data Catalog Vocabulary	The use of annotation properties to express metadata of datasets
Disease Ontology	The use of classes, concepts and properties to add domain knowledge about diseases
Friend of a Friend	The use of concepts and classes to express organizations
GeoNames	The use of entities and classes to express geographical information
Web Ontology Language	The use of concepts, classes and properties to develop ontology
RDF RDFS	The use of concepts, properties and classes to develop RDF triples
Statistical Data and Metadata eXchange	The use of concepts, classes and properties as dimensions, coded properties and measures to express dataset fields
Semnatic Web Rule Language	The use of rules to add complexity to knowledge graph
Extensible Markup Language XML Schema Defintion	The use of concepts, datatypes and properties to develop ontology

Table 1:Re-used Ontologies

whereas the property qb:sliceKey links the slice key to the data structure definition. The Observation Groups differ from slices in their domain. While slices can only contain data from a particular dataset, observation groups have no such condition.

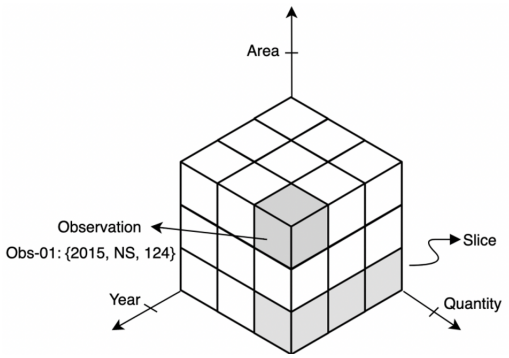

Fig. 2. RDF Cube Conversion[12]

The observations, instances of class qb:Observation, are the collection of dimension, measure and attribute values. Each observation, traditionally, includes the values for all the measures, called the multi-measure approach. However, the W3C also recommends using

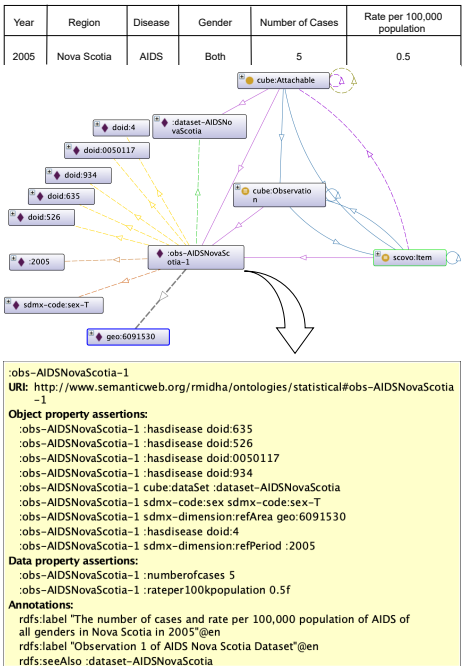

Fig. 3. Sample Observation.

a single measure approach wherein the observations are split up such that they contain the value of only one measure. This calls for the addition of a measure dimension which identifies the measure that that particular observation split contains the value for. Due to its complicated implementation, expense on space-time efficacy and minimal case-use benefit, the study drops the single measure approach in favor of the multi-measure approach. The observations are attached to the dataset by the qb:observation property and the respective slices by the qb:observationGroup property.

The ontology design described above can be used to incorporate datasets from independent and diverse sources. It is thus able to function as a generic primer to convert datasets to rdf cubes and link statistical data. To bring this concept home, the datasets chosen for this study were selected from different sources. The ontology design was found to be compatible with the open statistical datasets regardless of sources.

4.1.2. Interlinking: Disease Datasets

The datasets for diseases covered statistical data on diseases such as AIDS, HIV, Hepatitis, Typhoid and Tuberculosis. The datasets contain the number of cases and rate per 100,000 population of a certain disease for a given gender, time and region. These datasets were interlinked using a common data structure definition resulting in dimensional links with reason and logic.

Also, since diseases are a separate branch of study and form an exhaustive taxonomy, the ontology was linked to the disease ontology (DOID) [4]. The use of Geonames [5] to represent regional dimension information instead of literals adds another possibility for knowledge inference and creation. This allowed the addition of semantics to statistical data.

Also, AIDS data from Alberta was linked to data from Nova Scotia. This is a demonstrative example to show the ease of integration of external datasets in this ontology possible due to the strict adherence to principles and standards as per the W3C and RDF Cube modelling.

4.2. Questions

The web portal can be used to obtain information about datasets. However, due to lack of semantic knowledge and linked datasets, it cannot answer complex questions. Through the knowledge graph, the aim is to demonstrate how adding semantics to link datasets can help infer and create knowledge. A case study using disease datasets was developed to enable this process. The questions asked of the datasets are :

Question 1: What diseases by infectious agents occur in Nova Scotia?

Question 2: What viral infectious diseases have occurred in Nova Scotia in a given year?

Question 3: What is the maximum number of HIV cases reported in Nova Scotia?

Question 4: Which type of bacterial/viral infection is most common in Canada?

4.3. Rules

The use of SPARQL queries to obtain the following can be complex and even unattainable. Various rules can be applied to the ontology to add another layer of semantics thus enriching the knowledge base. Semantic Web Rule Language (SWRL) is an example of a Rule Markup Language which is an initiative to standardize the publishing and sharing of inference rules. It is built upon XML, RDF, and OWL. The syntax is based on the OWL concepts of classes, individuals and properties. It is more expressive but it is undecidable. The rules that were created to incorporate existing ontologies with this ontology to build semantic relationships. The Semantic Query Web Rule Language (SQWRL) is an extension of SWRL to achieve granularity and complexity in results by querying rules without needing to infer axioms. The rules are defined as :

- The relationship concerning diseases lacks transitive property. This implies that if a person has disease x which is a form of disease y, the graph cannot infer that person x has disease y implicitly. The rule states that,

$$stat : hasdisease(?x, y) \wedge doid : is\_a(?y, ?z) \implies stat : hasdisease(?x, ?z)$$

- The rule explicitly defines the transitive nature of doid:is_a property. The rule states that,

$$doid : is\_a(?x, ?y) \wedge doid : is\_a(?y, ?z) \implies doid : is\_a(?x, ?z)$$

- The highest number of cases of a particular infection recorded can be obtained using rule,

$$stat : numberofcases(?x, ?n) \implies sqwrl : max(?n)$$

  The rule can be made highly specific by using constraints on type of disease, region and period as well as gender information.
- The most common infections can be interpreted as infections with number of cases higher than a given threshold can be stated as

$$stat : numberofcases(?x, ?y) \wedge swrlb : greaterThan(?y, N) \implies sqwrl : select(?x, ?y)$$

The rule can be made highly specific by using constraints on threshold N serving as cut-off to classify common infections as well as other dimensions such as region, period, gender and disease.

4.4. Queries

We used the built-in SPARQL tab in Protege to pose the queries against the knowledge base through additional semantics which cannot be explicitly expressed through linkage. The queries can be also im-

[4] https://disease-ontology.org

[5] https://www.geonames.org

plemented using SQWRL tab built-in to Protege to retrieve SWRL rules' enriched data, without having to import the axioms.

Query 1: The query aims to answer the first question regarding diseases by infectious agents in Nova Scotia.

```
SELECT  DISTINCT ?Disease
WHERE { ?subject stat:hasdisease ?object; sdmx-dimension:refArea ?a.
        ?object rdfs:label ?Disease; doid:is_a doid:0050117.
        ?a rdfs:label "Nova Scotia"@en}
```

Disease
"tuberculosis"@en
"primary bacterial infectious disease"@en
"bacterial infectious disease"@en
"viral infectious disease"@en
"human immunodeficiency virus infectious disease"@en
"hepatitis B"@en
"typhoid fever"@en
"acquired immunodeficiency syndrome"@en

Fig. 4. Query 1

Query 2: The query aims to answer the second question. It tries to elicit the viral infectious diseases that occurred in Nova Scotia in 2017.

```
SELECT  DISTINCT ?Disease
WHERE { ?subject stat:hasdisease ?object;
        sdmx-dimension:refArea ?a;sdmx-dimension:refPeriod ?p.
        ?object rdfs:label ?Disease; doid:is_a doid:934.
        ?a rdfs:label "Nova Scotia"@en.?p rdfs:label "2017"@en}
```

Disease
"human immunodeficiency virus infectious disease"@en
"acquired immunodeficiency syndrome"@en

Fig. 5. Query 2

Query 3: The query aims to answer the third question. It elicits the maximum number of cases of HIV reported in Nova Scotia.

```
stat:numberofcases(?x, ?n) ^ stat:hasdisease(?x, doid:526) ^
sdmx-dimension:refArea(?x, geo:6091530) -> sqwrl:max(?n)
```

max(?n)
21

Fig. 6. Query 3

Query 4: The query aims to answer the fourth question. The information about common infections is retrieved. The threshold is taken such that the number of cases are greater than 10.

The results of all queries were cross-checked and validated for accuracy and completeness. The records returned are correct and all possible records are returned.

```
stat:hasdisease(?x, doid:0050117) ^
stat:numberofcases(?x, ?n) ^ swrlb:greaterThan(?n, 10)
-> sqwrl:select(?x)
```

x
stat:obs-HIVNovaScotia-1
stat:obs-HIVAlberta-1

Fig. 7. Query 4

4.5. Quality Check

Knowledge graphs on the Web are a backbone of many information systems that require access to structured knowledge, be it domain-specific or domain- independent. Knowledge graph refinement methods can differ along different dimensions [5]. The objective of knowledge graph refinement is to enhance the overall quality of the knowledge graph. It includes identifying and subsequently adding the missing knowledge as well as correction of erroneous information. The metrics to determine the quality of a knowledge graph have been theorized based on the various refinement techniques. The metrics are utilized as per the nature of the ontology and knowledge graph generated. As this study focuses on developing a knowledge graph for linked statistical open data, it calls for a horizontal ontology and domain-independent knowledge graph.

Quality Check	Description	Metric	Value
Accuracy	The correctness and validity of the information presented, verified against a legitimate source.	Spelling Error Rate	0%
Domain-specificity	A horizontal or shallow ontology (high) covers more domains but not in-depth and a vertical or deep ontology (low) domain specific.	Inheritance Richness	77%
Consistency	The adherence to a structure i.e. precision.	Inconsistent Terms Ratio	0%
Informative	The information conveyed by ontology on the basis of relationships.	Relationship Richness	64%

Table 2:Quality Check Metrics With Values

The basic metrics for determining the quality are accuracy and precision which are also applicable in the case of this study. The ultimate litmus-test for a knowledge graph is the information that is conveyed by it. For generation of a good knowledge graph, it is important that apart from being accurate and precise, it conveys a lot of information and also has potential and room for improvement.To determine some of these metrics, the tool OntoMetrics [6] has been utilized.

[6]https://ontometrics.informatik.uni-rostock.de/ontologymetrics/

5. Discussion

The study was motivated by the lack of meaning behind open statistical data. The generation of a knowledge graph for open statistical data was the primary objective of this study. The meticulous use of standards by W3C and their correct implementation yielded a robust graph that was able to infer and create knowledge. The datasets, primarily belonging to the field of health and medicine, exhibited inter-linking through structure, semantic relationships through integration of external ontologies and logical inference and creation of knowledge because of rules. This is supported and demonstrated by the complex questions pertaining to diseases answered through SPARQL queries. In doing so, the graph linked datasets from different portals and regions, of different diseases and to external ontologies. The results of the queries in previous section were cross-validated for precision and completeness manually against the source data. This exercise reaffirmed the quality check metrics that pointed to the knowledge graph being highly accurate and precise.

The methodology for creating a knowledge graph for open statistical data consisted of a workflow starting from gathering data. The open data portal of Nova Scotia with more than 500 datasets was chosen for this study and a set of Health and Wellness dataset were structured according to a design ontology. The ontology designed to develop the graph adheres to best practices and standards thereby allowing for expansion, modification and flexible re-use. The ontology is domain independent and able to accommodate cross-dimensional open statistical data. The development of a vocabulary and subsequently an ontology requires one to capture the logic behind data. This often requires a deep understanding of the domain. This lack of domain expertise is mitigated through use of external ontologies. The problem of scalability and flexibility of knowledge graphs is identified as a key challenge in a number of current studies. These issues were kept in mind during planning and execution of ontology design as well as the entire workflow. This ensures efficient data retrieval in the early stages of the graph with a small network as well as in the later stages when the network, owing to its flexibility, grows larger in size due to addition of diverse datasets.

The enrichment of data through inter-linking faced a challenge due to lack of web resources. The open data repositories of government portals are lacking which is detrimental to research in the field of semantic web development. The integration of semantics through rules in the SWRL tab of Protege posed certain issues as well. The limited debugging tools and documentation on the web coupled with scarcely descriptive error logs led to challenges in fixing run-time errors. Also, technological gaps in the software due to compatibility issues with certain versions caused difficulties.

The implementation of the knowledge graph poses a unique challenge of readability. The future use in research and final reach to the end-user is only feasible if the graph contains entities with adequately rich metadata. Furthermore,the average end-user may not be aware of the formal semantics and may not possess the domain knowledge for datasets but may need to access the data. The ontology had to not just link to complex concepts that made it conceptually correct at the fundamental level but also be easy-to-understand at the highest abstraction level. The possible lapses in readability due to language barriers, inadequate domain expertise and scarcely populated metadata is a big problem. To mitigate this, standard metadata vocabularies were implemented, external ontologies to supplement domain expertise were utilised and entities were derived from globally compatible resources.

6. Conclusion

The study demonstrates that the integration of open statistical data from multiple sources using ontologies and interlinking features of semantic web to generate a knowledge graph enables the performance of advanced data analytics. The ontology designed to develop the graph adheres to best practices and is domain-independent thus allowing expansion, modification and flexible re-use. The study also demonstrates the innumerable use cases made possible through semantic enrichment i.e data with meaning can be linked and queried in infinitely many ways. As a proof of concept, the study also uses said graph to answer complicated questions which the open statistical data portal in its current state cannot answer owing to the lack of semantics and meaning.

However, there are some shortcomings. The study faces the issue of inadequate populated classes due to technical gaps in the built-in axiom tools as well as source data. The various limitations and inadequacies of the research in the field of semantic web in regard to tools has already been identified. The knowledge base enrichment is made effective through inclusion of data from diverse sources which has proven to be a pitfall as open statistical data portals are sparse. The

best way forward is to develop semantic web sources and complimentary open data so as to drive forward the research in the field. This is possible through mainstream adoption and commercial implementation to make available the benefits of this field to global users.

7. Acknowledgement

The work conducted in the study has been funded by the MITACS Research Training (IT21970) and NSERC (Natural Sciences and Engineering Research Council) Discovery Grant (RGPIN-2020-05869).

References

[1] J. Marden, C. Li-Madeo, N. Whysel, and J. Edelstein, "Linked open data for cultural heritage: Evolution of an information technology," in *SIGDOC 2013 - Proceedings of the 31st ACM International Conference on Design of Communication*, 2013.

[2] R. P. Lourenço, "An analytvis of open government portals: A perspective of transparency for accountability," *Government information quarterly*, vol. 32, 2015.

[3] A. Gregory and M. Vardigan, "The Web of Linked Data: Realizing the Potential for the Social Sciences," tech. rep., 2010.

[4] E. Kalampokis, D. Zeginis, and K. Tarabanis, "On modeling linked open statistical data," *Journal of Web Semantics*, 2019.

[5] H. Paulheim, "Knowledge graph refinement: A survey of approaches and evaluation methods," *Semantic Web*, 2017.

[6] B. Stvilia, "A model for ontology quality evaluation," *First Monday*, vol. 12, no. 12, 2007.

[7] D. Vrandečić, *Ontology Evaluation*, pp. 293–313. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009.

[8] M. Alobaidi, K. M. Malik, and S. Sabra, "Linked open data-based framework for automatic biomedical ontology generation," *BMC bioinformatics*, 2018.

[9] Z. Syed, A. Pädia, T. Finin, L. Mathews, and A. Joshi, "UCO: A Unified Cybersecurity Ontology," in *AAAI Workshop - Technical Report*, 2016.

[10] M. N. Asim, M. Wasim, M. U. G. Khan, W. Mahmood, and H. M. Abbasi, "A survey of ontology learning techniques and applications," 2018.

[11] R. E. Cyganiak and D. E. Reynolds, "The RDF Data Cube Vocabulary," *W3C Recommendation*, 2014.

[12] E. Rajabi, "Towards linked open government data in Canada," *International Journal of Metadata, Semantics and Ontologies*, vol. 14, no. 3, pp. 209–217, 2021.

[13] C. van Ooijen, B. Ubaldi, and B. Welby, "A data-driven public sector: Enabling the strategic use of data for productive, inclusive and trustworthy governance," OECD Working Papers on Public Governance 33, OECD Publishing, May 2019.

[14] P. Escobar, G. Candela, J. Trujillo, M. Marco-Such, and J. Peral, "Adding value to Linked Open Data using a multidimensional model approach based on the RDF Data Cube vocabulary," *Computer Standards and Interfaces*, 2020.

[15] C. Debruyne, D. Lewis, and D. O'Sullivan, "Generating executable mappings from RDF data cube data structure definitions," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2018.

[16] J. Klímek, J. Kučera, M. Nečaský, and D. Chlapek, "Publication and usage of official Czech pension statistics Linked Open Data," *Journal of Web Semantics*, 2018.

[17] Z. Zhang, P. Yu, H. C. Chang, S. K. Lau, C. Tao, N. Wang, M. Yin, and C. Deng, "Developing an ontology for representing the domain knowledge specific to non-pharmacological treatment for agitation in dementia," *Alzheimer's and Dementia: Translational Research and Clinical Interventions*, 2020.

[18] V. Dornauer, M. Ghalandari, K. Höffner, F. Jahn, B. Schneider, A. Winter, and E. Ammenwerth, "Challenges and solutions while developing hito, a health it ontology," 07 2019.

[19] E. Katis, H. Kondylakis, G. Agathangelos, and K. Vassilakis, "Developing an ontology for curriculum and syllabus," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2018.

[20] P. Chen, Y. Lu, V. W. Zheng, X. Chen, and B. Yang, "Knowedu: A system to construct knowledge graph for education," *IEEE Access*, vol. 6, pp. 31553–31563, 2018.

[21] J. Xu, S. Kim, M. Song, M. Jeong, D. Kim, J. Kang, J. F. Rousseau, X. Li, W. Xu, V. I. Torvik, Y. Bu, C. Chen, I. A. Ebeid, D. Li, and Y. Ding, "Building a PubMed knowledge graph," *Scientific Data*, vol. 7, no. 1, p. 205, 2020.

[22] M. Koho, E. Hyvönen, E. Heino, J. Tuominen, P. Leskinen, and E. Mäkelä, "Linked Death—Representing, Publishing, and Using Second World War Death Records as Linked Open Data," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2017.

[23] A. Hofmann, S. Perchani, J. Portisch, S. Hertling, and H. Paulheim, "DBkWik: Towards knowledge graph creation from thousands of wikis," in *CEUR Workshop Proceedings*, 2017.

[24] A. Callahan, J. Cruz-Toledo, and M. Dumontier, "Ontology-Based Querying with Bio2RDF's Linked Open Data," *Journal of Biomedical Semantics*, 2013.

[25] M. Moshref and R. Al-Sayyad, "Developing Ontology Approach Using Software Tool to Improve Data Visualization (Case Study: Computer Network)," *International Journal of Modern Education and Computer Science*, 2019.

[26] M. R. Kamdar and M. A. Musen, "PhLeGrA: Graph analytics in pharmacology over the web of life sciences linked open data," in *26th International World Wide Web Conference, WWW 2017*, 2017.

[27] M. Mohd Ali, R. Rai, J. N. Otte, and B. Smith, "A product life cycle ontology for additive manufacturing," *Computers in Industry*, 2019.

[28] E. Ammenwerth, V. Dornauer, M. Ghalandari, F. Jahn, N. De Keizer, and A. Winter, "An ontology for describing health IT interventions: Methodological considerations," in *Studies in Health Technology and Informatics*, 2019.

[29] P. Jain, P. Hitzler, A. P. Sheth, K. Verma, and P. Z. Yeh, "Ontology alignment for linked open data," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2010.

[30] S. Staab, M. Erdmann, A. Maedche, and S. Decker, "An Extensible Approach for Modeling Ontologies in RDF(S)," in *Knowledge Media in Healthcare*, 2011.

[31] S. Decker, S. Melnik, F. Van Harmelen, D. Fensel, M. Klein, J. Broekstra, M. Erdmann, and I. Horrocks, "Semantic Web: The roles of XML and RDF," *IEEE Internet Computing*, 2000.

[32] I. Kollia, B. Glimm, and I. Horrocks, "SPARQL query answering over OWL ontologies," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2011.

[33] J. B. Lamy, "Owlready: Ontology-oriented programming in Python with automatic classification and high level constructs for biomedical ontologies," *Artificial Intelligence in Medicine*, 2017.

[34] K. Höffner, J. Lehmann, and R. Usbeck, "CubeQA—question answering on RDF data cubes," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2016.

[35] J. Zheng, M. R. Harris, A. M. Masci, Y. Lin, A. Hero, B. Smith, and Y. He, "The Ontology of Biological and Clinical Statistics (OBCS) for standardized and reproducible statistical analysis," *Journal of Biomedical Semantics*, 2016.

[36] M. Atzori, G. M. Mazzeo, and C. Zaniolo, "Querying RDF Data Cubes through Natural Language [Discussion Paper]," in *CEUR Workshop Proceedings*, 2018.

[37] B. Lantow, "Ontometrics: Putting metrics into use for ontology evaluation," in *Proceedings of the International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, IC3K 2016, (Setubal, PRT), p. 186–191, SCITEPRESS - Science and Technology Publications, Lda, 2016.