The Netherlands' Climate Knowledge Graph

Giorgia Nidia Carranza Tejada i6231006

March 2020

1 Introduction

An essential difficulty of this decade is the consequences of global warming. The topic is highly discussed, given the alarm of the scientific community about the outcomes that may come.

Related to the theme, the Netherlands Climate Graph intends to be a knowledge graph focused on collecting the data corresponding to the causes and consequences of climate change as the geographical information.

This instrument aspires to become a means of facilitating scientific research. The graph will be populated through different official sources, and then the environmental information will be connected to their localization. Finally, the completed project will allow us to query the system to retrieve knowledge.

2 Problem

A relevant topic these days is climate change. The problem of global warming is generated by several factors such as the greenhouse effect since the daily activities maximize the effect, provoking the earth's heat to rise even further and long-lasting changes in all components of the climate system[2]. Also, other reasons are deforestation, destruction of marine ecosystems, and population increase[9]. Besides, the consequences of climate change include a change in ecosystems and desertification, rising sea level, acidification of the oceans, extreme weather phenomena, etc. This project aims to collect data relative to both the causes and consequences of climate change relative to the Netherlands and answer to the following research questions:

- Which are the places in the Netherlands, which have experienced the largest increase in rains between 1970 to 2017?
- What is the difference in fossil consumption between 1970 to 2014 in the Netherlands?

3 Significance

The project was created to provide a unique tool that collects all the data relevant to study the climate.

At the moment, several projects and institutes give open access to distinct types of information. However, the data can be fragmentary and incomplete to have a complete view of the current climate situation. As an example, for this project, various sources were used to retrieve the knowledge.

This tool can be a useful instrument for the research purpose, and to demonstrate the actual problem that the Earth is undergoing.

4 Related Work

As already nominated, the issue of climate change is widely discussed by the scientific community. The relevance of the topic can be noticed by the number of publications about global warming, which are about 2 million in Google Scholar's search engine. Also, the organization Intergovernmental Panel on Climate Change (IPCC)[6] provides policymakers with regular scientific assessments on climate change.

Because of the significance of the problem, it would be expected to easily find the appropriate data for scientific research in a single verified tool. However, this is not the case, since the data are fragmented, covering only a part of the knowledge, and spread through different official sources. Also, the approach of knowledge graph related to the environment is rarely discussed. The paper [11] discusses a similar approach to develop an end-to-end automated methodology for incrementally constructing Knowledge Graph for Earth Science.

5 Innovation

As previously explained, the climate change is a relevant topic of this century. The topic, as explained in the section Related Works, is widely explored by the scientific community. Therefore, this project doesn't aim to realize experimental research on a particular relation between causes and consequences implied by the global warming. Several papers, researchers, and panels studied this area for years to find a correlation between parameters aiming to prevent the damage inflicted on Earth.

Instead, the innovative approach proposed by the paper consists of the methodologies used to process the information in a Knowledge Graph. It was chosen these type of instrument since it helps the integration of knowledge from different areas through shared property or vocabulary, as well as, generate links between entities from datasets of a different source. The last evidence was particularly connected to one of the reasons for the project, because, even if there were a wide availability of data connected to the climate, the information was distributed in different sources depending on the scientific parameter.

Besides, the project focused on the collection of data climate information about the Netherlands. However, the structure of the Knowledge Graph allows the map expansion to a global scale.

6 Methodology

The elements used in this work are discussed in the following section. In detail, in the next subsection will be described the sources and the data used. Following, the data modeling to the map of a graph will be explained. To conclude, it will be determined the entities connected through LIMES.

6.1 Data Source

This paper intends to collect the data relative causes and consequences of climate change in a single software. The types of information are of various field, therefore, the origin of them does not come from a single resource.

The National Centers for Environmental Information[3] provides access to archives of oceanic, atmospheric and geophysical data. From this source were obtained meteorological information of stations in different cities of the Netherland covered in the years between 1951 and 2018. The meteorological data includes the number of days with higher or equal to 0.01 inch/0.254 millimeters in the year or to 0.1 inch/2.54 millimeters. Also, it includes the latitude and longitude of the station.

The Statistics Netherlands[12] provides reliable statistical information and data. The dataset used presents the air quality from different sources, in a period between 1990 to 2018. The atmosphere condition includes information about carbon dioxide, non-methane volatile organic compounds, methane, sulfur dioxide, dinitrogen oxide, carbon monoxide, nitrogen oxides, ammonia and particle matter in a unit of million kgs. Besides, the sources and the periods are univocally identified by a key. So, the meaning of each identifier is included in two different datasets.

DataHub gives access to curate, create and find dataset. The project chosen in this case is CO2 Emissions from Fossil Fuels since 1751 By Nation[4] , which has as a source the Carbon Dioxide Information Analysis Center (CDIAC)[1]. It contains information about the fuel consumption by country between 1751 to 2014. In detail, it includes the quantity of carbon emission from solid fuel, liquid fuel, and gas fuel, cement production, and gas flaring.

The project world-cities[14] from Github offer the dataset derived from Geonames. The information provided includes the name of the city, the relative country and the geonameid that univocally identified the city. However, the dataset does not include the geographical details about the size of the population, the longitude, the latitude, and the country ISO. The knowledge is directly provided by Geonames, which provides dumps of places for each country. However, in this case, it was considered only sites of the Netherlands.

The databases listed cover geographical and scientific information. The data were provided in a CSV format and almost all the dataset were contained in one file, in exception to NOAA which provide 386 files.

6.2 Data model

The figure 1 shows the steps required to construct a knowledge graph from structured data.

The process starts by cleaning the data, because some of the files, in particular those derived from NOAA, were incorrectly formatted. Then, RDF Mapping Language (RML)[5] was used as the mapping language to convert from CSV to RDF. The instrument extends Relational Databases to RDF Mapping Language (R2RML) [10], which was used only for the conversions for Relational Databases. RML is a mapping language defined to express customized mapping rules from heterogeneous data structures and serializations to the RDF data model.

By this tool, it should define the structures of the triples to be included in the graph. In this case, the subjects used are eight and used shared vocabulary to define the types:

- Places, to describe entities with a physical or geographical extension. The items come from the Geonames and, they are of type schema:Place, which expresses the definition more appropriately.
- The city defined by the Github's project used the type schema:City.
- The entity country is defined by the DBpedia's class Country. It represents a distinct region in geography.
- The data originated from NOAA are mapped by using the subject Weather Station, which corresponds to the element Q190107 of Wikidata.
- The fossil fuel formed by natural processes such as anaerobic decomposition of buried dead organisms is the subject of the data relative to DataHub's project. It resembles the class of the element Q12748 of Wikidata.
- The air quality collected by CBS is correlated to the class Q56245086 of Wikidata. The definition of the entity that is chemical, physical, biological, and radiological characteristics of air, appropriately resembles the entity.
- The sources correlated to CBS's data are connected to the type pmlp:Source, which defines the sources of information.
- The period also connected to CBS's data is correlated to the class gleifbase:Period.

Following, through Linked Open Vocabulary[8], it was defined properties by using shared vocabularies.

In this case, for the terms associated with geographical entities, sources, and periods, most of the predicates reuse commonly defined terms. Instead, for the proprieties related to the chemical composition of the air, the carbon fuel emissions and meteorological, were not contained among the existing terms.

In the case of the predicates defined by NOAA, it was reused the same terminology defined in the original files since they are very specific terms related to the precipitations:

DP01 Number of days with ¿= 0.01 inch/0.254 millimeter in the year. Values originally recorded in inches as 0.01 are stored as 3 tenths of a millimeter.

DP10 Number of days with ¿= 0.1 inch/2.54 millimeter in the year. Values originally recorded in inches as 0.10 are stored as 25 tenths of a millimeter.

In the other case, the terminology reuse Wikidata terms that exactly match their definition.

The tool RMLmapper [13] was used to automatically convert the CSV file to an RDF Mapping Language with the parameters previously explained. The complete structure of the Knowledge Graph is shown in the figure 2.

6.3 Interlinking

Some connections through the different datasets were established by the tool LIMES [7]. The instrument discover the similarity based on heuristic between the entities by comparing one or more predicates. In this case, the links were established between:

- The subjects' Station and Place, since the Meteorological station gives the information about its localization. So, the tool connects the predicate localization of Station to the label of the City.
- The subjects' Place and Station have the same interconnection as the previous point, but the map is inverse.
- The subjects' Fossil fuel and Country, since the information of the first element refers to a nation. The comparison is between vcard:country-name of Fossil fuel and rdfs:label of Country.
- The subjects' City and Country have the connection that each city is part of a country. Therefore, the tool combines vcard:country-name of City and rdfs:label of Country.

  The metric used to realize interconnection is the function trigrams with a threshold of 0.99 for the links accepted instead of 0.6 to review.

7 Results

The resulted Knowledge Graph combines information relative to the geographical position of places, cities, and countries, as well as scientific and meteorological information, for example, the air quality and the emission of carbon fossil. The classes defined in the subsection 6.2 appears in the class hierarchy section of GraphDB, as in the figure 3.

And similarly, the interconnections establish in the subsections, to enhance the environmental information to have also a geographical background, are shown by the figure 4.

The following Knowledge Graph, consent also to retrieve climate information to answer the initial previous research questions. The SPARQL questions appear in Appendix A. The first query results the places with the largest increases of rains between 1970 to 2017 are Woltersum, Hoogwoud, Laren, Marken, Markenesse, Obdam, Laaghalen, Formerum, Herbayum and SteenWijksmoer. Instead, the difference in fossil consumption between 1970 to 2014 is 6918 metric tons of carbon.

8 Conclusions

The Knowledge Graph of the Netherlands' climate change intends to be an innovative instrument collecting data relative to the environment and the geography. Since the information was spread through different platforms, one of the difficulties was to retrieve the correct data. Therefore, the dataset used derived from official sources.

Another objective was interconnecting the environmental information to geographical data. The phase was realized through the tool LIMES.

The last purpose, the research questions initially proposed were answered in the section Results.

In conclusion, in the future, the project can be extended to global knowledge relative to all the countries. The Knowledge Graph is already compact to geographical information, therefore it could be enhanced by environmental data.

References

[1] *Carbon Dioxide Information Analysis Center (CDIAC) [Online]*. Available: `https://cdiac.ess-dive.lbl.gov/`. 2017 (accessed March, 2020).

[2] Climate Change. *IPCC Synthesis Report. Summary for policymakers 2014*. 2014.

[3] *Climate Data Online (CDO) - The National Climatic Data Center's (NCDC) Climate Data Online (CDO) provides free access to NCDC's archieve of historical weather and climate data addition to station history information [Online]*. Available: https://www.ncdc.noaa.gov/cdo-web/. 2011 (accessed March, 2020).

[4] *CO2 Emissions from Fossil Fuels since 1751, By Nation [Online]*. Available: https://datahub.io/core/co2-fossil-by-nation. 2018 (accessed March, 2020).

[5] Anastasia Dimou et al. "RML: a generic language for integrated RDF mappings of heterogeneous data". In: (2014).

[6] *IPCC - Intergovernmental Panel on Climate Change [Online]*. Available: https://www.ipcc.ch/. 2019 (accessed March, 2020).

[7] *LIMES [Online]*. Available: https://github.com/RMLio/rmlmapper-java. 2010 (accessed March, 2020).

[8] *Linked Open Vocabularies (LOV) [Online]*. Available: https://lov.linkeddata.es/dataset/lov/. 2011 (accessed March, 2020).

[9] *NASA: Climate Change and Global Warming [Online]*. Available: https://climate.nasa.gov/. 2020 (accessed March, 2020).

[10] *R2RML: RDB to RDF Mapping Language [Online]*. Available: https://www.w3.org/TR/r2rml. 2012 (accessed March, 2020).

[11] R. Ramachandran et al. "Building Scalable Knowledge Graphs for Earth Science". In: *AGU Fall Meeting Abstracts*. Vol. 2017. Dec. 2017, IN33B–0110.

[12] *Statistics Netherlands (CBS) [Online]*. Available: https://www.cbs.nl/. 1899 (accessed March, 2020).

[13] *The RMLMapper [Online]*. Available: https://github.com/RMLio/rmlmapper-java. 2019 (accessed March, 2020).

[14] *World cities [Online]*. Available: https://github.com/datasets/world-cities. 2016 (accessed March, 2020).

A SPARQL queries

First SPARQL query:

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX wikidata: <https://www.wikidata.org/wiki/>
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX time: <http://www.w3.org/2006/time#>
PREFIX noaa: <https://www.ncdc.noaa.gov/cdo-web/>

SELECT ?label ?location ?rainDiff WHERE  {
```

```
     { SELECT ?label1 ?rain1
        WHERE  {
            ?station1 a  wikidata:Q190107;
                      rdfs:label ? ?label1;
                      time:year ?year;
                      noaa:dp01 ?rain1;
                      noaa:dp10 ?rain101.
            BIND ( (?rain011 + ?rain101) as ?rain1)
            FILTER(?year = 2017)

        }}
        ?station a  wikidata:Q190107;
          time:year ?year;
          rdfs:label ? ?label;
          dbo:location ?location;
          noaa:dp01 ?rain;
          noaa:dp10 ?rain10.
    BIND ( (?rain01 + ?rain10) as ?rain)
    FILTER(?year = 1970)
    FILTER(?label = ?label1)
    BIND ( (?rain1 - ?rain) as ?rainDiff)

} ORDER BY DESC(?rainDiff)
limit 10
```

Listing 1: First SPARQL query

Second SPARQL query:

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX wikidata: <https://www.wikidata.org/wiki/>
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX time: <http://www.w3.org/2006/time#>
PREFIX vcard: <http://www.w3.org/2006/vcard/ns#>

SELECT ?country ?fuelDiff WHERE {
     { SELECT ?label1 ?fuel1
        WHERE {
            ?fuel2017 a  wikidata:Q12748;
                      vcard:country-name ? ?label1;
                      time:year ?year;
                      dbo:fuelConsumption ?fuel1.
            Filter(?year = 2014)
            Filter(?label1 = "NETHERLANDS")

        }}
```

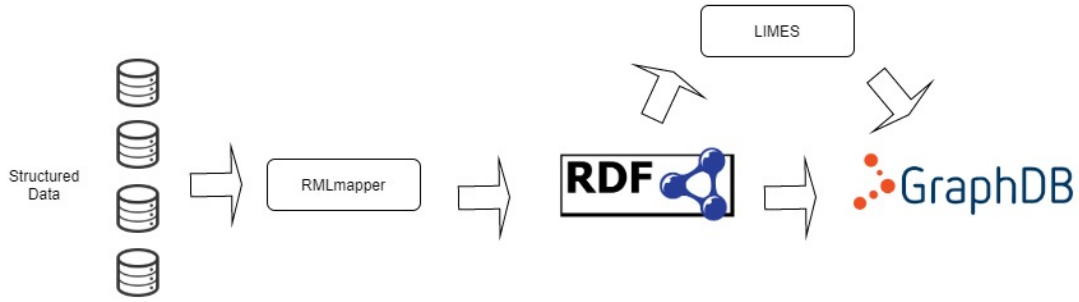

Figure 1: The structure of the process to obtain the Knowledge Graph

```
        ?station a  wikidata:Q12748;
          time:year ?year;
          vcard:country-name ?country;
          dbo:fuelConsumption ?fuel.
    Filter(?year = 1970)
    Filter(?country = "NETHERLANDS")
    bind ( (?fuel1 - ?fuel) as ?fuelDiff)
}
```

Listing 2: First SPARQL query

B Figures

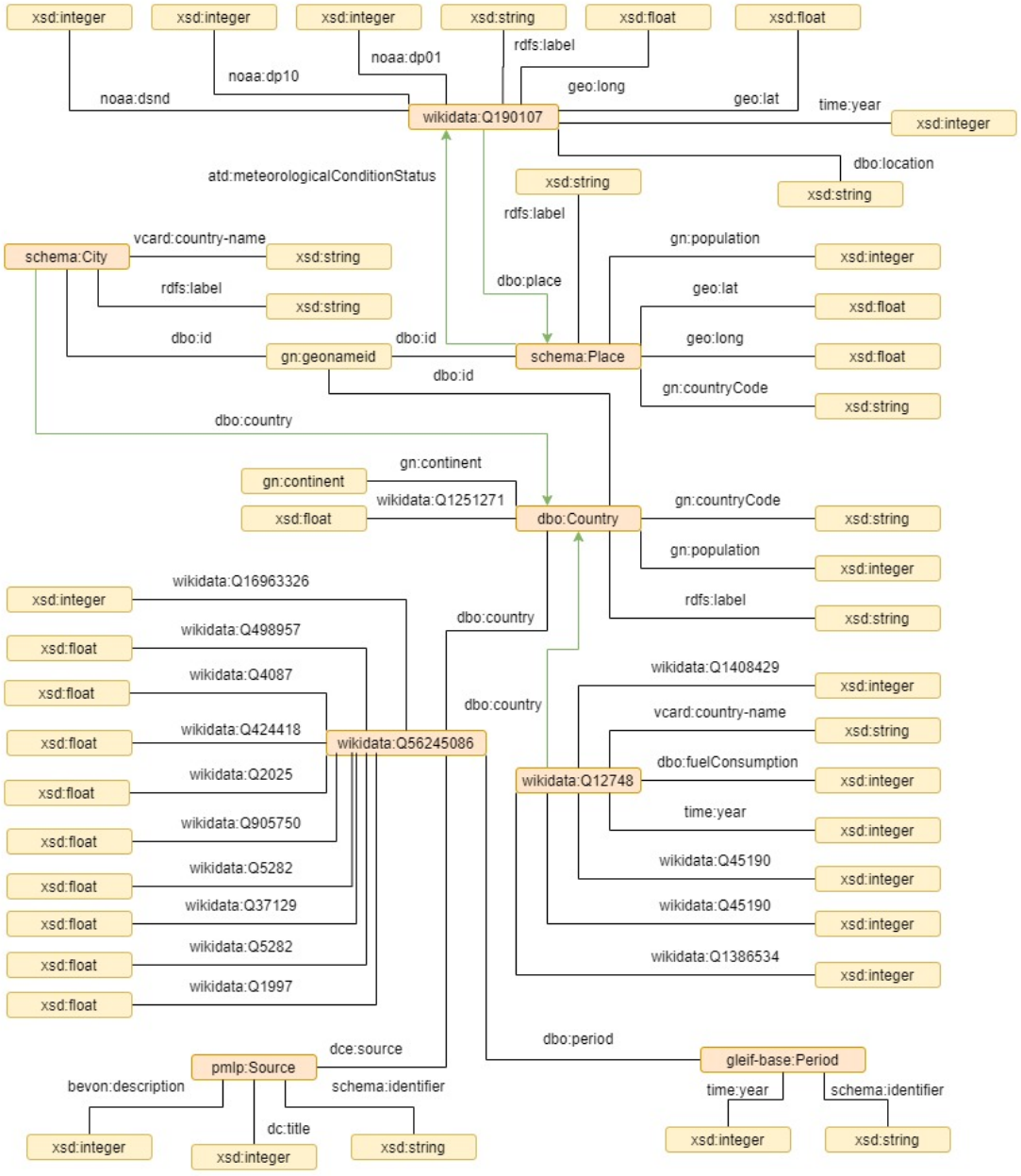

Figure 2: Map of the Knowledge Graph

Figure 3: The classes of Knowledge Graph

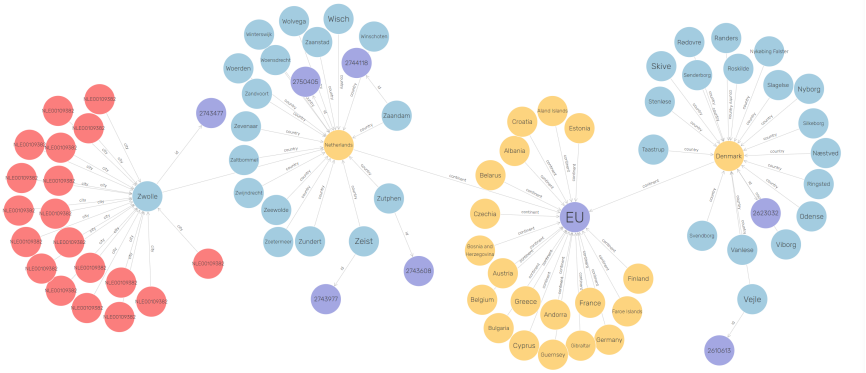

Figure 4: Visual Graph of GraphDB