A Proposal for Semantic Integration of Crime Data in Mexico City

Francisco Carrillo-Brenes[1](✉), Luis M. Vilches-Blázquez[2], and Félix Mata[1]

[1] Instituto Politécnico Nacional, Unidad Profesional Interdisciplinaria en Ingeniería y Tecnologías Avanzadas, Mexico City, Mexico
f.carrillo.brenes@gmail.com, mmatar@ipn.mx

[2] Instituto Politécnico Nacional, Centro de Investigación en Computación, Mexico City, Mexico
lmvilches@cic.ipn.mx

Abstract. Crime is a common problem in big cities where the authorities regularly update data crime reports. In Mexico City, the crime reports are available as open data. However, other relevant data are not connected to them (e.g., socioeconomic data). Therefore, the socioeconomic and geographic data can help understand how the crime is characterized and what social indicators are related to it. In this research, we explore how data crime reports are described and how they can be associated in an Ontology with other data, such as socioeconomic and geographic data. The goal is to discover the social indicators related to a particular crime in a specific area by using SPARQL queries from a knowledge representation. Then, data sets from crime reports, socioeconomic and geographic data from 2016 were integrated to explore crime behavior in Mexico City. The work uses a NeOn methodology in which resources from existing ontologies or non-ontological resources can be mixed. Next, a set of SPARQL queries is defined to extract the knowledge from ontology and discover the associations between crime in geographic and socioeconomic domains. The results showed a set of queries where it is possible to know where a crime occurred and what other factors are associated with the crime and help to identify possible patterns among them.

Keywords: Crime data · Mixed data · Behavior crime data · Ontology

1 Introduction

Crime occurrence is a phenomenon that affects the geographic area and its inhabitants. Impacting the status of the area, the living of the inhabitants, and how they make decisions. Today, data crime incidence is generated and published as open data in Mexico City. Institutions and organizations use this information to treat and understand the phenomena from different perspectives.

The traditional approach to generating and storing the crime reports consists of schemas of data (files and databases), limiting the capability to retrieve the relevant information, identifying associations, or combining the information with other data sources to reveal trends and patterns. Thus, a knowledge representation model (ontology) can

M. F. Mata-Rivera et al. (Eds.): GIS LATAM 2020, CCIS 1276, pp. 30–48, 2020.
https://doi.org/10.1007/978-3-030-59872-3_3

be built to represent the crime reports; it could improve the access and retrieve the relevant data to identify the associations and patterns. It generates that the original dataset's interoperability is increased.

In this work, data from three domains are integrated socioeconomic, geographic, and crime to reveal associations. Crime reports and socioeconomic data can be transformed into knowledge, represented using the knowledge representation model [1]. The ontology is the model used in this work, which is defined as a formal, explicit specification of a shared conceptualization [2]. It represents formal knowledge as a set of concepts of a domain using a vocabulary. This vocabulary is used to declare the types, properties, and relations between the concepts.

The dataset of crime reports from Mexico City is described as an ontology; it was developed based on scenario-based methodology. These scenarios emphasize the join, re-engineering, and reuse of ontological and non-ontological resources. Ontology is built using Protégé [available at https://protege.stanford.edu/], using the resources to define the concepts of the dataset, the properties, and relations.

The ontology can be used to develop an RDF file, where the dataset is described as semantic triples which is represented in the form subject-predicate-object expressions. Also, this RDF was published in a server, in order to make queries and retrieve data from the file. These queries were made using a particular language, which is SPARQL. This language is used to retrieve and also RDF management, like data creation, modification, and erase.

The rest of the paper is organized as follows: related work is presented in Sect. 2, while in Sect. 3, the datasets used are described. Section 4 the development methodology used in this work, and Sect. 5 shows the conclusions and future work.

2 Related Work

This section presents some works related to using an ontology to describe crime data and its analysis.

In [3], Jalil et al. used violent crimes as crime variables. The software TopBraid Composer was used to develop the ontology, defining relations and attributes. The ontology was used to implement a prototype named CrimeAnalysis, in which data about a new case is introduced and matches the preexisting data in the model. In the current work for the ontology development was used Protégé and for the RDF was used Python. When a query is made in this research, socioeconomic data related to an area where a crime occurred, is retrieved.

In [4], the data source were newspaper articles. The author obtained entities from the articles using natural language processing techniques. The ontology was developed using the entities extracted, and they added other existing ontologies. In the current work, the data source is obtained from an open data repository.

While in [5], a new ontology was developed named SMONT. Its primary purpose was to solve crimes. Its data source were complaints found in social networks, and these were extracted with natural language processing. Besides, the authors used open data repositories as a data source. The ontology was developed with the NeOn methodology, and the software used is Protégé. They used the ontology as a search tool for a kind of crime with SPARQL queries. The current work is similar, both use crime reports as a data source, but in the current work, socioeconomic data was added as an additional data source. The ontology development was made following the NeOn methodology and with Protégé software, and with the queries is possible to retrieve spatial-temporal characteristics and the socioeconomic characteristics.

In [6], the authors used the crime reports from social media and news. From there, they extracted entities using NLP techniques. With entities extracted and defined, the possible relations between them were recognized and declared in the ontology development. They developed the ontology using TopBraid. The current work uses official crime reports. The recognition of entities is made in the official data.

In [7], an existing ontology was adapted to the crime domain. The entities added to this ontology about crime were extracted from documents using natural language processing, specifically with an algorithm named SVO (Subject, Verb, Object) to analyze sentences to construct triples. In the current work, the ontology is made from scratch, reusing some concepts from other ontologies, and the entities are defined from the dataset.

While in [8], the authors made an extension to [5]. They used official crime reports and extractions from social networks. Subsequently, they extended the ontology SMONT with new classes extracted from the database compressing the crime reports from police and social networks. They built a knowledge base and applied machine learning to identify patterns according to a crime classification. In the current work, there will also be a union of the crime reports and the extraction from social networks, but in the part of classification and pattern detection, deep learning will be used.

In [9], the authors developed an ontological knowledge base reusing an event ontology for criminal events and causes. The events were extracted from news headlines using NLP techniques. Criminal events are extracted from the official crime reports from the city, and the ontology development reuses some existing ontologies in the current research.

Another ontology to describe criminal events was developed in [10]. The authors extracted news related to criminal events from several web pages and with NLP techniques extracted the entities and separated the entities per crime. With these entities, they developed the ontology to describe the selected crimes. In contrast, the current work used official crime reports from an open data repository, and the entities were extracted from there, besides the use of socioeconomic data to relate the crime reports with the social indicators.

In Korea, Gun-woo et al. [11] developed an ontology, based on information extracted from official reports and unstructured data regarding intrusion theft, and they implemented an ontology-based search service. In the current work, the ontology was developed from the official reports, and it was added socioeconomic indicators to find a relation between them. The publication in Virtuoso allows performing queries of a specific crime and retrieve indicators related.

3 Data Sources and Wrangling

Data used in this research consist of two datasets:

1. The first dataset is the investigation folders of the attorney general's office of the city of Mexico (available at http://datos.cdmx.gob.mx/explore/dataset/carpetas-de-investigacion-pgj-de-la-ciudad-de-mexico/). Its license is CC BY, which allows the user to share and adapt the dataset. It contains reports made by crime victims in many dependencies in Mexico City. In the crimes reported, 292 different crimes occurred in the 16 mayoralties in the city and some municipalities of Mexico's state.
2. The second dataset is the social backwardness index, which is issued by, institute for measuring the social development in Mexico (CONEVAL by acronym in spanish) available at https://www.coneval.org.mx/Medicion/IRS/Paginas/Indice_Rezago_Social_2015.aspx). The license of the dataset is CC BY. It contains indicators as total population, illiteracy rate, percentage of people not attending the school, and percentage of people with primary education incomplete, among others.

The data wrangling process [12] was performed into both datasets. It consists of extracting, cleaning, and integrating with other datasets. Moreover, the identification and datatypes of each useful variable are achieved.

The crime dataset contains 292 crimes, some crimes with a low number of reports. Then, we selected the crimes with more than 2000 reports, obtaining only 27 crimes as follows: confidence abuse, sexual abuse, threats, third party property damage to an automobile, intentional property damage, facts complain, dispossession, documents forgery, fraud, injuries in a collision, intentional injuries, drug dealing, home theft without violence, violent business theft, nonviolent business theft, vehicle driver theft, subway passenger nonviolent theft, cellphone violent theft, cellphone nonviolent theft, violent theft on a public road, vehicle accessories theft, object theft, theft of objects from inside a vehicle, violent vehicle theft, vehicle nonviolent theft, identity usurpation, and family violence.

In addition, were maintained only the reports concerning Mexico City, i.e., the 16 mayoralties of Mexico City. The dataset resulting had 144082 tuples for the year 2016. It has 15 variables, including the datatype containing the coordinates where crime reported occurred. Eleven variables are text type. These variables include occurrence month, crime reported, crime type, prosecution, agency, research unity, mayoralty where the crime occurred, suburb, streets, and a Geopoint, which is the union of the coordinates. There is a variable with date datatype, which is when the crime occurred, and the last variable refers to the year.

The second dataset picked, the social backwardness index includes data about people. This dataset has values for five different years, 2000, 2005, 2010, 2015, and 2020. In this work, the data of 2015 was maintained, and the other removed because the crime data is of 2016. There are five variables in the dataset; all of them have a numeric datatype, specifically, float datatype. The variables are the number of people living in an area, the illiteracy rate, the percentage of people between 6 and 14 years old not attending the school, and the percentage of people without access to health services, and the percentage of people 15 years older with primary education incomplete.

Both datasets, the crime reports, were integrated into one dataset using Python and Pandas library. The social data was added in new columns per each mayoralty found in the crime dataset, resulting in a dataset with 20 variables and the 144082 tuples.

4 Methodology

The methodology consists of three phases: 1) building an ontology to describe crime and socioeconomic data following the NeOn [13] methodology. 2) RDF publication and queries construction to knowledge extraction of ontology 3) Identifying patterns from the results obtained by queries. It is represented in Fig. 1.

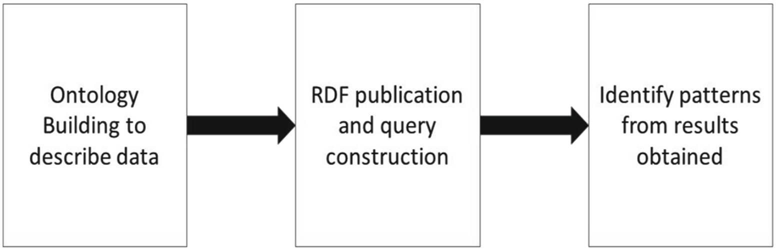

Ontology Building to describe data
RDF publication and query construction
Identify patterns from results obtained

Fig. 1. Methodology

4.1 NeOn Methodology

NeOn methodology is used to build the ontology. It considers the existence of ontologies related, the collaborative development of ontologies, the dynamic dimension, the reuse, and the re-engineering of the knowledge resources. Nine scenarios compose it, which a brief description of each one is the following:

Scenario one: "from specification to implementation," refers to ontology development from scratch, where the ontology requirements are specified: scope, the possible end-users, the dataset specifications, and competency questions. Besides, possible ontologies to reuse must be identified according to the datasets and the design of the URI-base, which is the identifier for the entire ontology.

Scenario 2, named reusing and re-engineering non-ontological resources, refers to deciding which resources can be reused in the ontology building and transforming those resources into ontologies. Scenario 3: Reusing ontological resources, refers to the use of existing ontologies, as a whole, as a module or as a statement.

In scenario 4, reusing and re-engineering ontological resources, existing ontologies can be reused and re-engineered before their integration in the new ontology. While in scenario 5 and 6 are similar called: reusing and merging ontological resources, the ontology developers reuse and merge existing ontologies to create a new ontology. In scenario 7: reusing ontology design patterns, it is designed patterns repositories to reduce modeling decisions. Scenario 8 "ontological resources" consists of restructuring existing ontologies and integrating them into the new ontology.

The scenario 9: localizing ontological resources, refers to translating the ontology developed in other languages to obtain a multilingual ontology.

These scenarios described can be combined in different ways.

4.2 Ontology Modeling

In this research, from NeOn [13] scenarios described in Subsect. 4.1, scenarios 1, 2, 3, 6, and 8 were used.

In scenario 1, the competency questions were defined. These questions are the ones that the ontology will answer; the questions must be according to the domain to describe. In our research, the domain is criminal incidence and socioeconomic data. Some of these questions are the following:

- How many crimes are registered?
- Where occurred a crime?
- When occurred a crime?
- Is there a socioeconomic indicator in a region?
- How many people live in a region?
- How many different crimes are there?
- What is the social indicator relating to a crime?
- What are the coordinates of a crime reported?
- Where there are more crimes reported?

The next step is to recognize the possible ontological and non-ontological resources used in ontology development. The ontological resources refer to ontologies already developed or some statements from ontologies regarding the development domain, which can describe the data and help answer the competency questions. Secondly, non-ontological resources refer to a knowledge resource that is not included or formalized in an ontology. In our work, it includes definitions of different crimes in crime reports dataset and definitions of some variables in the socioeconomic data and information of the region where the crimes occurred. The ontologies selected are shown in Table 1.

Table 1. Existing ontologies used.

Prefix	Meaning
QB	Vocabulary for data cube representation
QB4ST	Vocabulary for data cube representation with spatial temporal attributes
GEO	Geographic ontology
DBO	Dbpedia ontology

The explanation of each one element in Table 1 is the following:

- QB (available at https://www.w3.org/TR/vocab-data-cube/). Since the dataset is a relational database, this can be presented or modeled as an OLAP, and this ontology allows it to represent information from a table using the W3C RDF standard. It is focused purely on the publication of multi-dimensional data on the web.
- QB4ST (available at https://www.w3.org/TR/qb4st/). Is the ontology used to describe a data cube with attributes spatial-temporal. This ontology is an extension to QB ontology. It is used to define spatial-temporal attributes and measures.

- Geo (available at https://www.w3.org/2003/01/geo/). Since the data has longitude and latitude variables, the ontology used to represent coordinates was the geo vocabulary from W3C. This ontology is a *basic* RDF vocabulary that provides the Semantic Web community with a namespace for representing latitude, longitude, and other information about spatially-located things.
- DBO (available at http://dbpedia.org/). DBpedia resources were used to define some values of the socioeconomic data in the ontology. DBpedia is a project to export resources from Wikimedia to the semantic web, and it contains ontological resources.

The non-ontological resources were definitions of crimes, location information, and definitions regarding socioeconomic variables found in the social backwardness dataset. Moreover, the datasets used had metadata, which defines the names used in the datasets. Metadata was used to declare concepts in the ontology and establish relations between those concepts.

There are not ontologies to describe the crime in repositories. Thus, the classes and subclasses used to describe crime reports were defined using metadata. These classes include crime, crime category, prosecution, year, month, date, agency, location, population, and health services. The classes defined using metadata are shown in Table 2.

Table 2. Classes defined

Classes	Subclasses
Crime	
Category	
Prosecution	
Year	
Month	
Date	
Location	Mayoralty Suburb Street
Population	
Health services	
Agency	
Research unity	

Also, relations were defined using the dataset's metadata. The object properties defined with metadata and the data properties defined using the dataset include "hasIncompleteEducation" or "attendsSchool" are shown in Table 3.

For each property, a domain and a range had to be defined. Domain refers to the resource that has the property, and the range is the literal or resource that is affected by

Table 3. Object and data properties defined in the ontology

Object properties	Data properties
hasAgency	withoutHealth
hasCategory	attendsSchool
hasProsecution	incompleteEducation
hasDate	
hasLocation	
hasGeopoint	
hasResearchUnit	

the first resource. Most of the object properties defined have crime class as the domain and range different resources, while data properties defined has location as the domain.

Applying scenario 2, the non-ontological resources were picked, like crime definitions or places information, are integrated into the ontology or the dataset. Also, information about social indicators used was added and integrated to the original dataset.

In scenario 3, the ontological resources were picked. The existing ontologies added are shown in Table 1. These ontologies fit in the development, but not entirely. Besides, scenario six is also used, which refers, as mentioned, to reuse, re-engineering, and merge the ontological resources.

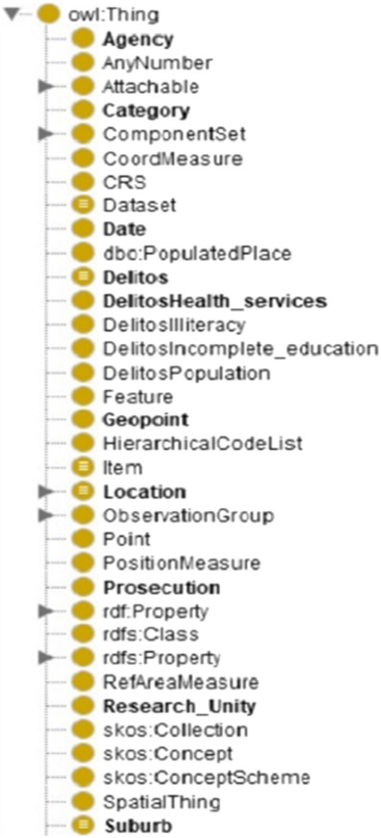
Now, applying scenario 6, the ontologies selected in scenario three (shown in Table 1) are modified to fit the new ontology and, subsequently, merge them with the new. The modifications were made in the conceptualization level. QB ontology was imported in this development entirely. As mentioned, a relational database can be described as an OLAP in the ontology. Thus, the entire classes and properties are imported into the new ontology.

In this, QB4ST is added too, also entirely, because the dataset used has attributes and measures spatial-temporal. In comparison, GEO ontology is imported entirely because this ontology has only classes to describe the coordinates.

The DBpedia ontology was imported, but this one has many resources and properties to use in an ontology. From this ontology, only one class was imported, which is "PopulatedPlace." This class was used to define the location class defined previously. Also, from this ontology, two data properties were used: population total, which describes the number of people living in a place, and this data property is used to relate location and populated place classes. Moreover, the other data property used was illiteracy rate, which refers to the percentage of people in a place with illiteracy; this property relates the location with a percentage value.

The last scenario picked, which is scenario 8 consists of three activities, modularization, pruning, and enrichment. In this ontology to describe the crime and social data, the prune was used to remove relations and attributes of the imported ontologies, especially from DBpedia ontology. Besides, two activities presented in the enrichment were made, the extension and specialization. In the first one, the ontology is extended with new concepts from the ontologies imported. In the second activity, the ontology is refined and specializes in specific concepts and relations.

All the classes defined with the ontologies imported are shown in Fig. 2.



owl:Thing
Agency
AnyNumber
Attachable
Category
ComponentSet
CoordMeasure
CRS
Dataset
Date
dbo:PopulatedPlace
Delitos
DelitosHealth_services
DelitosIlliteracy
DelitosIncomplete_education
DelitosPopulation
Feature
Geopoint
HierarchicalCodeList
Item
Location
ObservationGroup
Point
PositionMeasure
Prosecution
rdf:Property
rdfs:Class
rdfs:Property
RefAreaMeasure
Research_Unity
skos:Collection
skos:Concept
skos:ConceptScheme
SpatialThing
Suburb

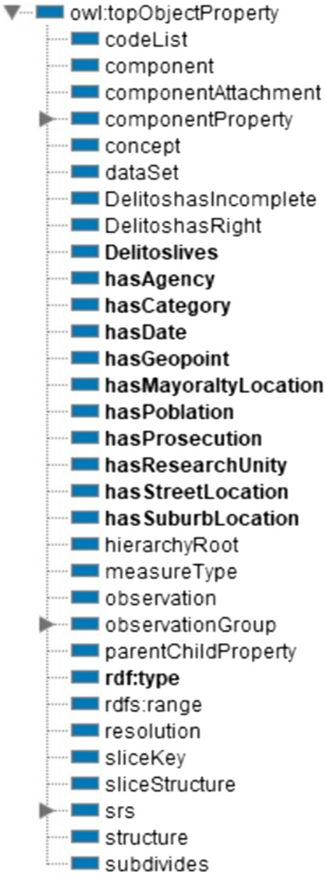
Fig. 2. Defined classes

While in Fig. 3 the entire object properties are shown.

The data properties defined to describe the dataset are shown in Fig. 4.

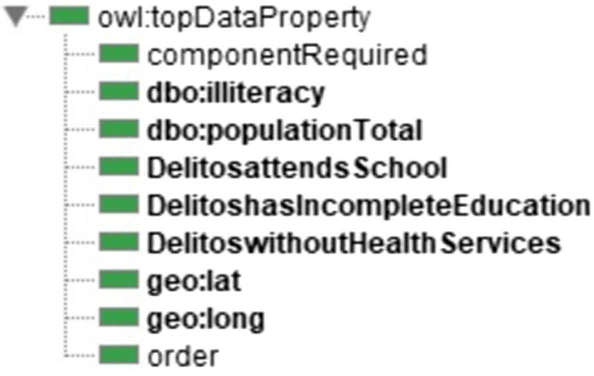
The crime class was defined with a set of axioms like something that always has a category, always has a date, always has a mayoralty location, on occasions has coordinates, on occasions has a second street.

In contrast, Location is always a populated place, a mayoralty, is divided into suburbs and streets. Besides, a location is always a populated place, has a population, has an illiteracy rate. The total axioms of the ontology, between the declared and the already declared in the ontologies imported, are 476 axioms.



owl:topObjectProperty
- codeList
- component
- componentAttachment
- componentProperty
- concept
- dataSet
- DelitoshasIncomplete
- DelitoshasRight
- **Delitoslives**
- **hasAgency**
- **hasCategory**
- **hasDate**
- **hasGeopoint**
- **hasMayoraltyLocation**
- **hasPoblation**
- **hasProsecution**
- **hasResearchUnity**
- **has StreetLocation**
- **has SuburbLocation**
- hierarchyRoot
- measureType
- observation
- observationGroup
- parentChildProperty
- **rdf:type**
- rdfs:range
- resolution
- sliceKey
- sliceStructure
- srs
- structure
- subdivides

Fig. 3. Defined object properties



owl:topDataProperty
- componentRequired
- **dbo:illiteracy**
- **dbo:populationTotal**
- **Delitosattends School**
- **DelitoshasIncompleteEducation**
- **DelitoswithoutHealth Services**
- **geo:lat**
- **geo:long**
- order

Fig. 4. Defined data properties

4.3 RDF Generation

Subsequently, an RDF file was developed, importing the ontology developed and the complete dataset containing the crime reports and the socioeconomic data. file. Using Python and the RDFlib library, an empty graph was defined, in which the triplets defining the data are added one by one.

The dataset was described as a data cube in the ontology developed. First, the definition as a datacube is added on the graph. At first, URI identifiers for each cube component were defined for the cube, dataset, cube properties, and components specifications. Next, the cube's measures and dimensions were defined. Definitions were made with classes in data cube ontology and with data cube for Spatio-temporal ontology.

Data related to the crime were defined as dimensions, the crime itself, prosecution, category, agency, and research unit. Also, location with its subclasses, i.e., the mayoralties, suburbs, and streets were defined as dimensions in the cube. The month of the report and the date of each report in the dataset were defined as temporal dimensions Using a data cube with Spatio-temporal data.

While the values about the socioeconomic like population, illiteracy, were defined as measures; the latitude and longitude coordinates were defined as spatial measures. In parallel, while the cube properties are being defined, the dataset is defined with the ontology developed, specifically with the classes and properties defined for crime reports.

Then, as the location of a crime is defined as a dimension property with the data cube part, it is also defined as a populated place, and his populations are described according to the data properties declared before. Crime is declared as a dimension property and also is declared as an instance of Crime class and the relations with the prosecution, the agency, the location, and its definition using isDefinedBy from RDF-schema.

With the declaration of each dimension, the next step was to define the data cube structure. Subsequently, the observations were defined, which is considered in the data cube ontology as the actual data, i.e., the observations are the tuples of the dataset. Each tuple is added in the graph, defining the number of the observation and appending all the data to its corresponding class of the ontology.

4.4 SPARQL Queries

The last part of this RDF development is the publication of the knowledge base, to make queries over the data in a semantic way with SPARQL language [14] using Virtuoso software [available at http://vos.openlinksw.com/owiki/wiki/VOS]. In this software, the RDF is uploaded, and with SPARQL, specific data in the RDF can be retrieved.

One query written in SPARQL language is shown below, in this, it was retrieved the mayoralty, the suburb and the population of each mayoralty.

```
Prefix nsl: <http://localhost:8890/cubos/carpetas/prop>
Prefix owl: <http://www.w3.org/2002/07/owl#>
Prefix qb: <http://purl.org/linked-data(cube#>
Prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
Select ?Mayoralty ?Suburb ?Population
Where
```

```
{
?obs rdf:type qb:Observation;
ns1:alcaldia ?Mayoralty;
ns1:colonia ?Suburb;
ns1:población ?Population.
}
Group by ?Mayoralty
```

The result of the query is a table showing the number of inhabitants per mayoralty and the suburbs per mayoralty. A fragment of the table obtained in the query is shown in Table 4.

Table 4. Population per mayoralty

Mayoralty	Suburb	Population
Benito Juárez	San José Insurgentes	391939
Álvaro Obregón	Tetelpan	746887
Azcapotzalco	Aldana	400708
Benito Juárez	Del Carmen	391939
Coyoacán	Barrio San Lucas	609627
Coyoacán	Prado Churubusco	609627
Coyoacán	Paseos de Taxqueña	609627
Álvaro Obregón	Central Camionera poniente	746887

In addition, other **Queries made in the RDF** are: Q1 = {Crime counting per mayoralty}, Q2 = {Crime counting per suburb}, Q3 = {Crime counting per crime}, Q4 = {Illiteracy rate}, Q5 = {Percentage of people with incomplete education}, Q6 = {Crime more reported in mayoralties and illiteracy rate}, Q7 = {Highest Illiteracy rate and the crime more reported} and Q8 = {Lowest illiteracy rate and crime with more reports}

Query Q1, allows to know the crimes per mayoralty in Mexico City, where Cuauhtemoc mayoralty had the highest numbers of reports with 23027 crimes reported in 2016. The next mayoralty is Iztapalapa with 21129 reports, followed by Gustavo A. Madero with 13359 and Benito Juaréz, where 13333 crimes were reported. In contrast, the mayoralty with the lowest number of crimes reported is Milpa Alta, which has 783 reports, preceded by La Magdalena Contreras with 2350 reports and Tlahuac with 3007 crimes reported. It is shown in the map of Fig. 5.

The zones with more crimes reported in Mexico city are colored in yellow and white, the crimes reported in other zones are colored in red, while the fewer crimes reported are colored in blue. As can be seen, all the mayoralties have suffered a crime.

In Q2, the crime reports per suburb distribution was retrieved. In the results, the suburbs with the highest number of reports were recognized. Some of them are in mayoralties

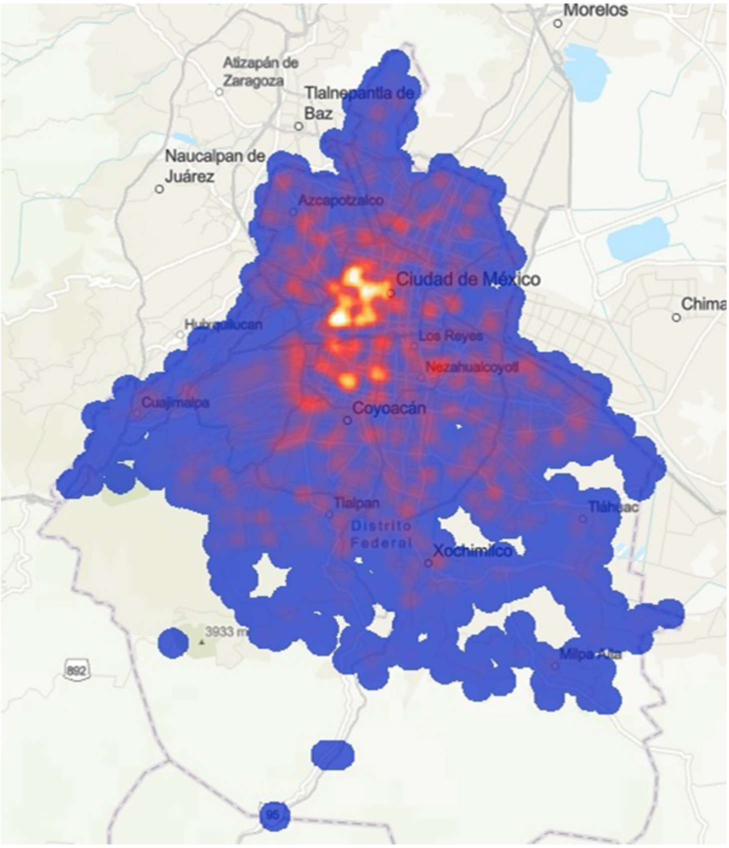


Fig. 5. Crime distribution (Color figure online)

with more crimes reported like Centro suburb, which had 4426 reports, or Doctores suburb with 2653 reports. Both located in Cuauhtémoc Mayoralty, which, as mentioned, is the mayoralty with the highest number of reports

The number of reports per crime was retrieved in query Q3, and the resulting table is shown in Table 5.

The crime with the highest reports in the data is domestic violence. At the same time, the crime with the lowest number of reports is subway passenger nonviolent theft with 620 reports.

Q4 was performed to retrieve the illiteracy rate of each mayoralty. The results show that Milpa Alta mayoralty is the one with the highest value, which is 3.31%, the second mayoralty is Iztapalapa with 2.37%. In contrast, the mayoralty with the lowest illiteracy rate is Benito Juárez, which value is 0.29, preceded by Miguel Hidalgo and Cuauhtémoc with 0.86 and 1.09, respectively. The results obtained were transformed into a map layer, shown in Fig. 6.

In which, each mayoralty is colored in a scale to indicate the region's illiteracy rate. The zones colored in red have a significant illiteracy rate. In comparison, the mayoralties

Table 5. Number of reports per crime

Crime reported	Number of reports
Domestic violence	17995
Objects theft	14925
Nonviolent business theft	13428
Facts complain	11099
Fraud	10587
Threats	9933
Violent theft on a public road	6290

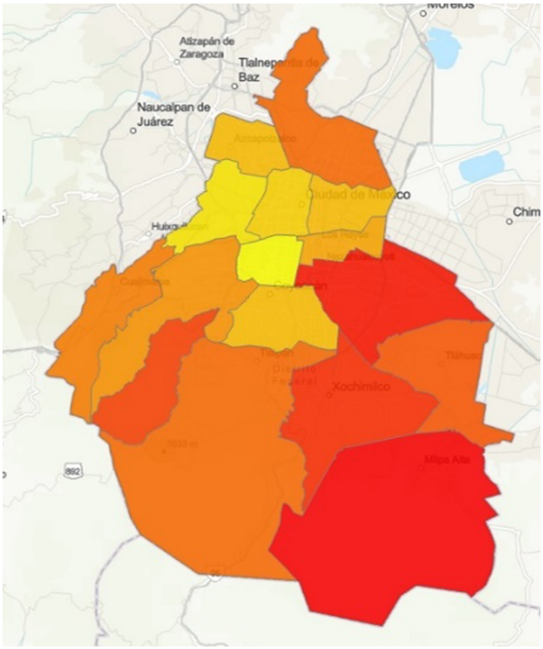


Fig. 6. Illiteracy rate

colored in yellow have a lower illiteracy rate. The illiteracy rate range is from 0.3 to 3.3%. At the same time, the crimes are represented as a heat map over the illiteracy rate

In order to retrieve another social indicator, Q5 was performed. In this query, the percentage of people with incomplete primary education. The results show that Milpa Alta, Iztapalapa, and La Magdalena Contreras are the three first mayoralties with a higher value, with 27.43%, 26.38%, and 25.33% respectively. In comparison, Benito Juárez is the mayoralty with the lowest percentage, which is 6.51%.

As the previous result related to social indicators with crime data, this table was transformed into a map layer, which is shown in Fig. 7.

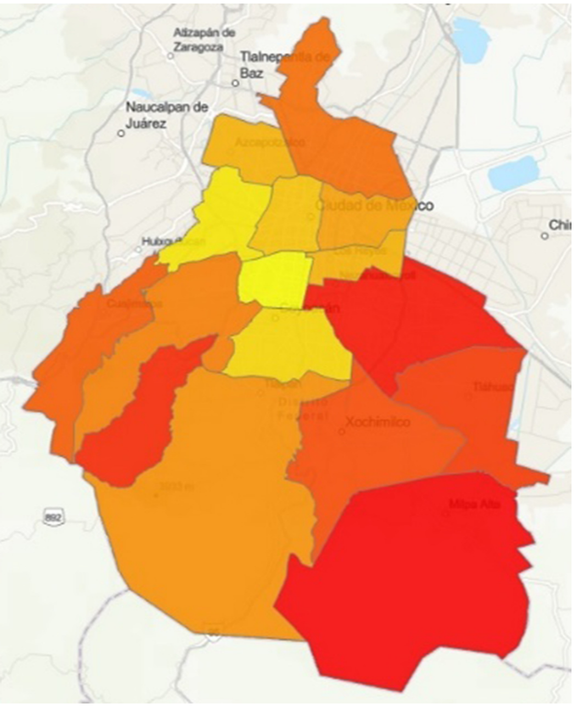


Fig. 7. Percentage of incomplete education

In order to relate the crime most reported in Mexico City and a social indicator Q6 was performed. In which, Iztapalapa, Gustavo A. Madero, and Tlalpan were retrieved as the mayoralties with more reports regarding domestic violence. Iztapalapa has 3837 reports; Gustavo A. Madero has 1879, and Tlalpan has 1372 reports. For these mayoralties, the illiteracy rate and domestic violence were retrieved. The illiteracy rate in Iztapapa is 2.37, in the second mayoralty is 1.76 and 1.84 for the third one. That result shows that the illiteracy rate is related to domestic violence. With an illiteracy rate of over 1.7%, these mayoralties have the highest numbers of domestic violence. Figure 8 shows the resulting map.

Performing Q7, the mayoralties with mayor illiteracy rates, and the number of domestic violence reports were retrieved. The mayoralties retrieved were Milpa Alta with

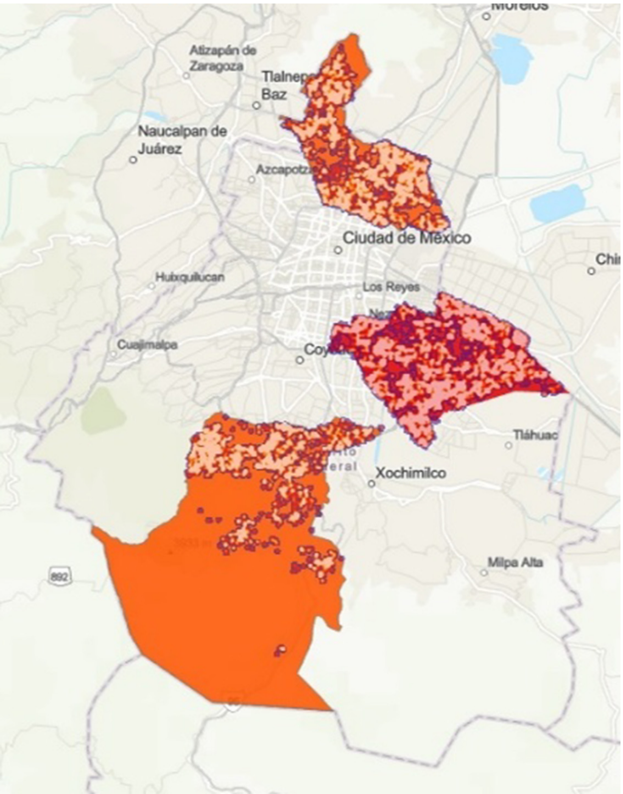


Fig. 8. Domestic violence and illiteracy rate

3.31%, Iztapalapa, which was already considered, Xochimilco with 2.28, and La Magdalena Contreras with 2.13%. Besides the number of domestic violence reports where the results show that Milpa Alta has 228, reports, Xochimilco 875, and La Magdalena Contreras 603. Compared to the previous result, the reports are not similar. However, in these mayoralties, there are fewer reports than in the previous result, and in these mayoralties the crime with the highest number of reports is domestic violence. The result was transformed, as previous results, into a map layer, shown in Fig. 9.

In order to compare the results, the Q8 query was performed. With this query, the three mayoralties with the lowest illiteracy rate were retrieved. The results show that Benito Juárez has 0.29%, Cuauhtemoc has 0.86%, and Miguel Hidalgo mayoralty has 1.09%. Besides, the domestic violence reports were retrieved, and the order in crime sorting according to the number of reports. In the three mayoralties mentioned is the fifth crime with more reports. Moreover, the number of domestic violence reports is 715 in Benito Juárez, 1489 in Cuauhtemoc, and 599 in Miguel Hidalgo. The results retrieved were transformed into a map layer shown in Fig. 10.

Even if domestic violence is the fifth crime reported in these mayoralties, the distribution in the heat map looks denser. The reason is that the number of crimes reported

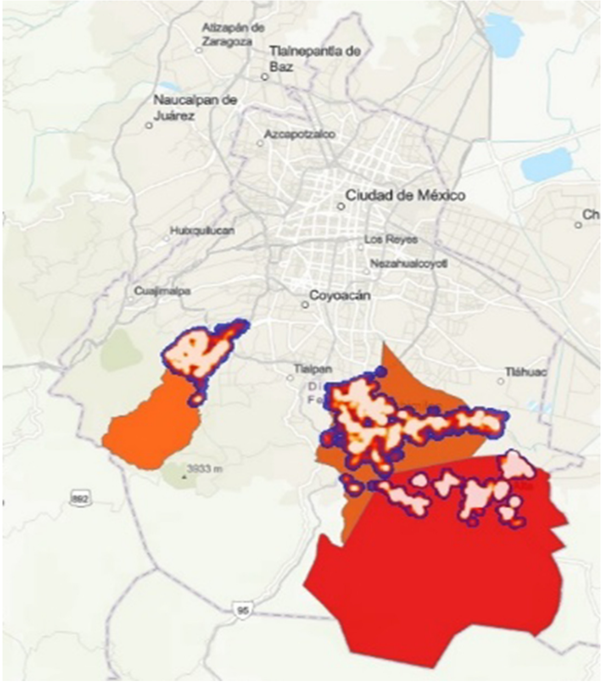


Fig. 9. Highest illiteracy rate and domestic violence

in the mayoralties shown in Fig. 10 is higher than the mayoralties mentioned in Q6 and Q7. Also, the urban zone is more significant in the mayoralties shown in Fig. 10.

According to results retrieved in Q6, Q7, and Q8, the illiteracy rate of over 1.7% can be considered a pattern in the domestic violence report. Performing more queries in data, in order to relate more social indicators with domestic violence in the areas mentioned, it will be possible to find other relations and consider them as patterns in the reporting of this crime. And not only with domestic violence, but with the other 26 crimes mentioned.

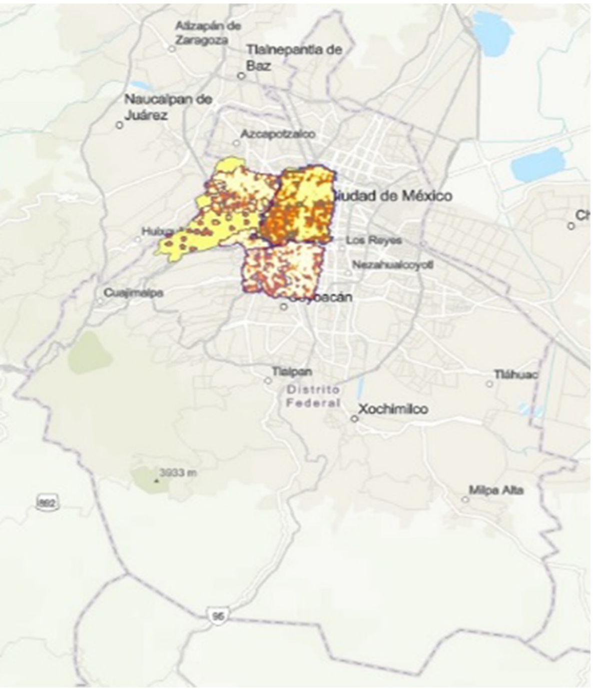


Fig. 10. Lowest illiteracy rates and domestic violence

5 Conclusions and Future Work

The presented work consisted of ontology development and an RDF generation to describe the crime reports in Mexico City and relate those reports with socioeconomic data from the area where the crimes occurred.

We presented the probable relations between the socioeconomic data and the crime reports. It was seen that some crimes have relations with some social variables, like relations between the illiteracy rate or the percentage of people with incomplete primary education with the number of crimes reported in an area. However, it is needed to make a probabilistic study to see if there is a correlation between the variables.

As future work, there will be added into the dataset new tuples using social networks extractions. Entity extraction will be made with natural language processing techniques.

Also, it will be considered new data about social indicators, is looking to add new about mobility and socioeconomic level from the mayoralties in the city. Once added, the correlation of these indicators with the crime reports will be calculated to decide the possible relevance of the study case indicators.

Although, as mentioned, the entities extracted can be added to the dataset of crime reports, and this dataset will be described in a new RDF using the ontology, which also could be enriched. However, another possibility is, once the RDF is published can be used queries to add new data to this knowledge representation model.

Instead of using ArcGIS, an RDF publisher with the possibility of visualization will be used to obtain the map at the same time. This publisher will avoid the data retrieved exportation and the map construction in ArcGIS.

It can be implemented machine learning techniques to identify the patterns and make the classification of each crime. In the part of the pattern, social data can be considered.

Machine learning techniques to identify the patterns and make the classification of each crime.

Also, this work will be published following the principles of linked data where anyone can access the SPARQL endpoint of the triple store developed in the RDF description of this dataset.

References

1. Pirnay-Dummer, P., Ifenthaler, D., Seel, N.M.: Knowledge representation. In: Seel, N.M. (ed.) Encyclopedia of the Sciences of Learning, pp. 101–211. Springer, Boston (2012). https://doi.org/10.1007/978-1-4419-1428-6
2. Gruber, T.: Toward principles for the design of ontologies used for knowledge sharing. Int. J. Hum. Comput. Stud. **43**, 907–928 (1993). https://doi.org/10.1006/ijhc.1995.1081
3. Jalil, M., Ling, C., Maizura, N., Mohd, F.: Knowledge representation model for crime analysis. Procedia Comput. Sci. **116**, 484–491 (2017). https://doi.org/10.1016/j.procs.2017.10
4. Thilagam, P., Srinivas, K.: Crime base: towards building a knowledge base for crime entities and their relationships from online newspapers. Inf. Process. Manage. **56** (2019). https://doi.org/10.1016/j.ipm.2019.102059
5. Kalemi, E., Domnori, E.: SMONT: an ontology for crime solving through social media. Int. J. Metadata Semant. Ontol. **12**, 71–81 (2019)
6. Acharya, G., Shakya, A.: Crime ontology extraction from news and social media. In: ICAEIC-2019, vol. 2, no. 1, Department of Electronics and Computer Engineering (2019)
7. Carnaz, G., Nogueira, V.B., Antunes, M.: Knowledge representation of crime-related events: a preliminary approach, Department of informatics, university of Évora, Portugal (2019)
8. Elezaj, O., Yayilgan, S.Y., Kalemi, E., Wendelberg, L., Abomhara, M., Ahmed, J.: Towards designing a knowledge graph-based framework for investigating and preventing crime on online social networks. In: Katsikas, S., Zorkadis, V. (eds.) e-Democracy 2019. CCIS, vol. 1111, pp. 181–195. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-37545-4_12
9. Reyes-Ortiz, J.A.: Criminal event ontology population and enrichment using patterns recognition from text. Int. J. Pattern Recogn. Artif. Intell. **33**(11) (2019). https://doi.org/10.1142/s0218001419400147
10. Rahma, F., et al.: Analysis and implementation of ontology based text classification on criminality digital news IOP Conf. Ser. Mater. Sci. Eng. **662**, 022135 (2019). https://doi.org/10.1088/1757-899x/662/2/022135
11. Ko, G.-W., Kim, S.-W., Park, S.-J., No, Y.-J., Choi, S.-P.: Implementation of ontology-based service by exploiting massive crime investigation records: focusing on intrusion theft. J. Korean Soc. Libr. Inf. Sci. **53**(1), 57–81 (2019). https://doi.org/10.4275/KSLIS.2019.53.1.057
12. Endel, F., Piringer, H.: Data wrangling: making data useful again. IFAC-PapersOnLines **48**, 111–112 (2015)
13. Suárez-Figueroa, M., Gómez-Pérez, A., Motta, E., Gangemi, A.: Ontology Engineering in A Networked World. Springer, Berlin (2014). https://doi.org/10.1007/978-3-642-24794-1
14. Zou, L.: SPARQL. In: Liu, L., Özsu, M.T. (eds.) Encyclopedia of Database Systems. Springer, New York (2018). https://doi.org/10.1007/978-1-4614-8265-9