International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XXXIX-B2, 2012
XXII ISPRS Congress, 25 August – 01 September 2012, Melbourne, Australia

NON-SPATIAL AND GEOSPATIAL SEMANTIC QUERY OF HEALTH INFORMATION

S. Gao [1], F. Anton [2], D. Mioc [2] and H. Boley [3]

[1] Department of Geodesy and Geomatics Engineering, University of New Brunswick, Fredericton, NB, Canada; E-Mail: sheng.gao@unb.ca
[2] National Space Institute, Technical University of Denmark, Denmark; E-Mail: fa@imm.dtu.dk, dmioc@space.dtu.dk;
[3] Institute for Information Technology, NRC, Fredericton, NB, Canada; E-Mail: Harold.Boley@nrc.gc.ca

Commission II – ICWG II/IV – Semantic Interoperability and Ontology for Geospatial Information

KEY WORDS: public health; ontologies; respiratory diseases; RuleML; geospatial data; semantic interoperability

ABSTRACT:

With the growing amount of health information and frequent outbreaks of diseases, the retrieval of health information is given more concern. Machine understanding of spatial information can improve the interpretation of health data semantics. Most of the current research focused on the non-spatial semantics of health data, using ontologies and rules. Utilizing the spatial component of health data can assist in the understanding of health phenomena. This research proposes a semantic health information query architecture that allows the incorporation of both non-spatial semantics and geospatial semantics in health information integration and retrieval.

1. INTRODUCTION

Health information systems are becoming increasingly important for public health security. Health data can be collected by hospitals, clinics, surveys, or any other health care facilities in different ways. The data collection process varies at different health organizations with different tools and methods. The integration of health data across service systems is a challenge (McLafferty, 2003). Indeed, health data are very heterogeneous and health standards have a wide variability in their implementation, and thus, one of the challenges is how to use information technology to enhance health information query and knowledge discover.

"The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation" (Berners-Lee, Hendler and Lassila, 2001). It can provide the semantic level of interoperability and facilitate the access to heterogeneous information sources. There are three sources of heterogeneity -- syntactic, schematic, and semantic -- that need to be considered (Bishr, 1998). Two types of semantic heterogeneity are distinguished (Lutz, Riedemann and Probst, 2003): one is cognitive heterogeneity that arises when two disciplines have different conceptualizations of real world facts; the other is naming heterogeneity which refers to different names for identical concepts of real world facts. Resolving semantic heterogeneity would greatly enhance the handling of syntactic heterogeneity and schematic heterogeneity (Bishr, Pundt and Ruther, 1999). Formal ontologies constitute an important notion of the Semantic Web. They are characterized as formal specifications of conceptualizations (Gruber, 1993). With well-designed ontologies, the meaning of distributed data can be unambiguously defined; semantic heterogeneity can be resolved; and therefore data sharing and integration can be enabled.

Considerable research has been conducted concerning the mapping and integration between different health ontologies (Lee, Supekar and Geller, 2006), (Pérez-Rey et *al.*, 2006), (Ryan, 2006).

Ontologies are usually expressed through the standard Web Ontology Language (OWL). DL (Baader et *al.*, 2003), which strives for decidability and usually for tractability, constitutes the formal underpinning for OWL deductive reasoning. DL represents knowledge through a TBox (terminology of concepts and properties) and an ABox (assertion of instances using the terminology). Rules, with Horn Logic as their formal underpinning, complement DL to express other kinds of knowledge in the Semantic Web (Grosof et *al.*, 2003).

To represent spatial relationships, the explicit storage or dynamic computation of spatial relationships is possible. Explicating all the possible spatial relationships between every two spatial objects is not necessary sometimes. The weakness of dynamic computation is the computation issues, while explicit storage leads to a significant storage and reliability issue (Jones et al., 2003). Klien and Lutz (2005) illustrated the definition of geospatial concepts based on spatial relations and automatic annotation of geospatial data based on a reference dataset. The annotation process uses DL in reasoning and focuses on the concept level. Smart et al. (2007) distinguished the multi-representation, implicit spatial relations, and spatial integrity characteristics of geospatial data, and claimed that rule expression for geo-ontologies needs to consider spatial reasoning rules and spatial integrity rules. Kammersell and Dean (2006) proposed the GeoSWRL, which is a set of geospatial SWRL built-ins. GeoSWRL allows users to include spatial relation operators in their query; however, the spatial data representation and processing abilities are not full integrated into the GeoSWRL.

2. GEOSPATIAL SEMANTICS

Geospatial semantics describe the underlying meaning of geopatial objects and their spatial relationships in the data. The spatial component of health data, which can show the geographical distribution of disease outbreaks, hospitals, clinics, air quality, and census, is of great importance in analyzing and visualizing health phenomena. Geospatial location provides a solution to link various sources. The spatial component in the

health data can be recorded implicitly or explicitly. Implicit spatial information treats the spatial attributes the same as non-spatial attributes, while explicit spatial information is special as it stores explicitly the geometric and/or topological information, including spatial reference and coordinate arrays. Geospatial location can be derived from implicit spatial information (e.g., the name of a location, Toronto), where corresponding knowledge is required for people or computers to understand its geospatial location. The use of explicit spatial information can support dynamic spatial relationship discovery and visualization of health data. Furthermore, new concepts and instances can be generated from existing health data with the use of explicit spatial information. For example, from the locations of infectious disease outbreaks, we can determine sensitive areas that are within certain distance from the disease outbreak location.

The enrichment of the health data with semantic metadata can enhance inference power in applications. Previous research generally handled geospatial location implicitly as text-based information (e.g., the name of a city) and defined their relationships using ontologies (e.g., a city is inside a province) in the health information query and integration. To relieve the efforts to explicitly define all the spatial relationships between spatial objects in health data integration, the consideration of geospatial semantics needs to be explored. Because of the advantages in supporting explicit representation of the spatial component, we endeavour to include explicit spatial information in the Semantic Web environment to utilize the geospatial semantics in health data.

The spatial component in the data stores the geometric and/or topological information. We can determine the spatial location or boundaries of the data from the geometric information. As the spatial relations exist between two spatial concepts or individuals, exploring the geospatial semantics using spatial relation can advance information query and discovery. Three types of major spatial relationships are topological, direction, and metrical relationship (Rashid et *al.*, 1998). Topological relationships relate to the concept of neighborhood; directional relationships require the existence of a vector space; and metric relationships require a distance. Topological relationships are invariable under continuous mappings while directional and metric relations may change during continuous mappings. The well-known formalism to reason about topological relationship in 2-dimensional space is the Nine Intersection Model (9IM), developed by Egenhofer, which considers boundaries, interiors, and complements intersection of two spatial objects (Egenhofer, 1991). Further improvement of this model is the Dimensionally Extended Nine Intersection Model (DE-9IM) that considers the 9IM of two spatial objects with the dimensions of -1 (no intersection), 0, 1, or 2 (Clementini and Di Felice, 1991), (Clementini and Di Felice, 1994). The commonly known topological predicates described by the DE-9IM include overlaps, touches, within, contains, crosses, intersects, equals, and disjoint. Plenty of research has been done on topological relationships between more complex spatial objects (Schneider, 2002), (McKenney et *al.*, 2007).

With spatial relationships present on many data sources, several applications or studies have been carried to capture geospatial semantics for facilitating data integration, query, and discovery. Perry et al. (2007) discussed the emerging field of extending semantic reasoning from the purely thematic dimension to the three dimensions: theme, space, and time. Kieler (2008) discussed the feasibility of identification of semantic relations between different onotlogies by exploring the geometric characteristics of the instances.

In addition, spatial operations can generate new spatial objects from existing spatial objects, such as spatial intersection and spatial union. As rules are suitable for describing concepts and relationships through complex property paths, it would be possible to represent spatial operations and spatial relationships of spatial objects as rules in the knowledge deduction. In this paper, we included the geometric representation in the Semantic Web, and applied ontologies and rules in health information reasoning and query. The respiratory disease information queries are used as examples in this study.

3. ARCHITECTURE FOR NON-SPATIAL AND GEOSPATIAL SEMANTIC QUERY HEALTH INFORMATION

Health concepts are related to non-spatial and geospatial attributes, as shown in Figure 1. The non-spatial attributes can be name, description, property, and time. Explicit representation of the geospatial attributes is about the geometry and the topology, which allows the discovery of geospatial semantics. Health concepts can be visualized with point, line, or polygon geometries that describe the spatial reference and coordinates of health data. Furthermore, health concepts can include cartographic attributes that specify the styles in map representation. For example, the non-spatial attributes of a health event can be event outbreak time, event type, and event description. The geospatial attributes of a health event can be point geometries showing the latitude and longitude of the event location. The cartographic attributes can describe the styles that are used to show the health event on maps. Relationships, including non-spatial and geospatial relationships, exist between the health concepts and health concept instances.

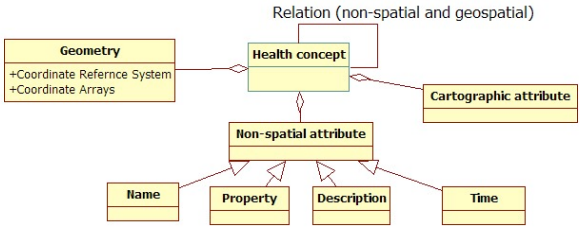

Figure 1. Health concept meta-model.

The architecture for semantic query of health information is shown in Figure 2. Health-related data can be accessed from various sources, such as files, database, Web Services, XML or Geographical Markup Language (GML). Various ontologies could exist in these data sources. These data sources are the essential content for the reasoning server, and will be translated to facts in the knowledge base of the reasoning server. The translation process can use methods like the Extensible Stylesheet Language Transformations (XSLT). The knowledge base in the reasoning server includes facts, ontologies, and rules. If different ontologies are used between data sources and the reasoning server, then ontology mapping is needed for translating between data sources. The ontologies are the formal representation of health concepts and their relationships in the non-spatial and geospatial dimensions. Rules, with the use of ontologies and facts, can deduce new health information. New concepts or knowledge can be described or deduced from rules, without the need to explicate all the knowledge in ontologies and facts. The application server is responsible for performing the business logic of applications (e.g., generating maps from corresponding health data). The query client is used to obtain health data or maps. User ontologies and templates can be designed in the query client for health data query. If user

ontologies are different from the ones at the reasoning server, ontology mapping will also be needed during the query process.

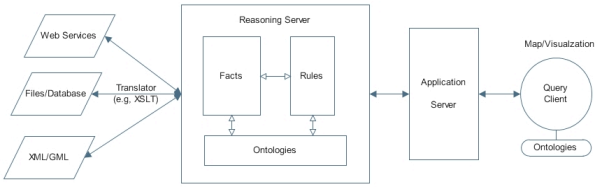

Figure 2. Architecture for semantic health information query.

4. ONTOLOGIES IN THE REASONING SERVER

The utilization of ontologies makes various concepts (e.g., subconcepts and superconcepts) connected. Depending on the requirements, different application ontologies can be created for health applications. To facilitate the health data exchange and query, a global ontology can improve the interoperability between the sources. Four types of ontologies are important in describing and visualizing health data: health domain ontologies, geometric ontologies, topological ontologies and cartographic ontologies. Health domain ontologies are defining health information models, concepts, and terminologies. Many standards exist in this field, such as HL7 standards, SNOMED-CT, and ICD-9. Geometric ontologies should be able to describe basic geometries types, such as point, line and polygon. Some attributes can be associated with the geometry types, including spatial reference and specific attributes (e.g., width for line strings). The European Petroleum Survey Group (EPSG) coordinate system codes are widely used in the exchange of geospatial data over the Internet. Spatial reference and coordinates arrays together can distinguish geometries. Topological ontologies describe a structural viewpoint on a domain and represent the connections between objects. Cartographic ontologies are related to the styles in map representation. Commonly, the point graphics, line graphics, polygon graphics, and chart graphics are useful for information visualization. For instance, the symbols of hospitals can be served as point graphics on maps to indicate the location of hospitals. The existence of health domain ontologies, geometric ontologies, and cartographic ontologies will form the basis for health concept and relation definition in application ontologies.

4.1. Rules in the reasoning server

Rules can be defined to deduce new information based on ontologies and facts. Besides the non-spatial attributes rules, geospatial rules are also applied in this framework. Although the definition of the geometric ontologies follows the same procedure as non-spatial ontologies, the inference of geometric relationships is different. The utilization of geometrics can add spatial analysis and cartographic representation functions into rules. Two types of rules are distinguished: reasoning rules and cartographic rules.

Reasoning rules are used to compare and deduce health information. They cover the semantic matching, spatial relationship operators, spatial operations, and cartographic comparison of the contents. These rules can be combined to form more complex rules.

RuleML (2012), which has co-evolved with SWRL (2004), SWSL-Rules (2005), WRL (2005), and RIF (2010), is the de facto open language standard for Web rules. Thus, RuleML is selected for the semantic retrieval of health information in our study. The RuleML's POSL presentation syntax is used as it is much less verbose than in the XML format. It is Prolog-like but also permits 'attribute->value' slots, as in F-logic. OO jDREW is open source and used in this study as the RuleML engine, because it supports RuleML's Naf Hornlog sublanguage, and backward/forward reasoning (OO jDREW, 2012). To use explicit spatial information in the reasoning process, the representation of spatial information in the RuleML engine is needed. Therefore, a geometry type is designed to support basic geometry types: point, linestring, polygon, multipoint, multilinestring, multipolygon, multimix. A point records the coordinate reference system (e.g., EPSG:4326 for the geographic coordinates in World Geodetic System 1984) and two-dimensional coordinates. Linestring records the coordinate reference system and a list of two-dimensional coordinate arrays for a line. A polygon can have a coordinate reference system, an out boundary, and many inner holes (inner boundaries). The out and inner boundaries record the coordinate arrays for the boundaries. Multipoint, multilinestring, and mulitpolygon can have one or more points, linestrings and polygons respectively. MultiMix contains collections of points, linestring, and polygons. With the declaration of geometries, the spatial operation (union, buffer, convexhull, difference, distance, intersection) and spatial relation operators (touches, contains, within, crosses, disjoint, equals, overlaps, intersects, covers, coveredby, iswithindistance) will be available for spatial reasoning in rules. Based on the design, a geometry type is added and a parser is created for parsing the geometries in OO jDREW. JTS Topology Suite is used in this study for spatial operation and relation operators. The JTS is an open source Java API for two dimensional spatial predicates and functions based on the DE-9IM model (VIVID SOLUTIONS, 2012). Several geospatial built-ins such as gpred_intersects, gpred_within, and gfunc_intersection are added into OO jDREW with the use of JTS library. For instance, the gpred_intersects built-in checks whether two geometries intersect or not; gpred_within built-in checks whether a geometry is inside another geometry or not; gfunc_intersection built-in is used to compute the intersection of two geometries.

5. RESULTS AND DISCUSSION

Respiratory disease information queries are used as examples in this study. The related data used in this study are collected from different organizations such as New Brunswick Lung Association, Service New Brunswick, Statistics Canada census, and Statistics Canada community health survey. Following the disease taxonomy of the respiratory diseases in ICD-9, an OWL-based taxonomy is built to describe different kinds of respiratory diseases and their relationships. A portion of the respiratory disease ontology is shown in Figure 3. The respiratory disease data are from hospital patient incidents, which record the admit time, three digital postcode, disease diagnosis category, age, and gender information. The geospatial location of three digital postcodes is explicitly specified with the center coordinates. In the health data collection, data may be collected at different spatial division of the real world. For example, patient incident data are collected through postcodes; the average income information from the Statistics Canada census is available in dissemination area or counties; the smoke rates from Statistics Canada community health survey are obtained in health regions.

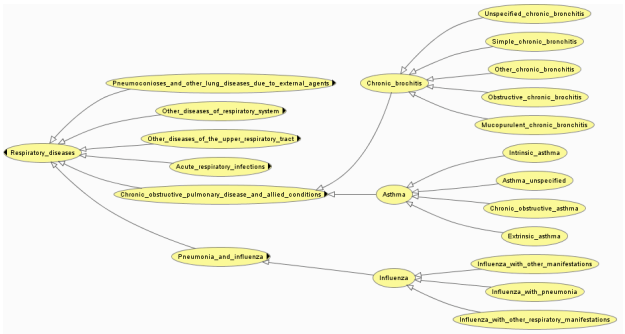

Figure 3. Fragment of the ontology on respiratory diseases

With these available data sources, the application ontologies used in the reasoning server are defined in our study. We created the entities: health event, hospital, health region, census division, three digital postcode, age, gender, smoke rate, and income. *Health event* can describe a variety of cases such as patient incidents, health training services, etc. The following properties associated with health events are considered here: the involved participants' age and gender, the admit time, disease category diagnosis, and the postcode. *Hospital* introduces the general information about hospitals, with attributes: name, address, city, province, telephone, fax, and geometry. *Health region* and *Census division* are two kinds of administrative boundaries. They have name, area, perimeter, and geometry attributes. *Postcode* describes the central location of the three digital postcodes. *Age* specifies the age type (e.g., senior, adult) as well as its age range. *Gender* defines male, female, and both types. *Smoke rate* and *income* show the value associated with the geometry name, age type, gender, statistic method, and year. In this study, the smoke rate data from Statistics Canada community health survey is collected at the health region level, and income information from Statistics Canada census is at the census division level. With the defined application ontologies, the data sources can be translated into the knowledge base of the reasoning server as facts. For example, a hospital admission of an 88-year-old inpatient diagnosed with "Influenza_with_pneumonia" on January 1st, 2000 is recorded as health_event(disease->?:Influenza_with_pneumonia;age->88:Integer;gender->Female;postcode->E1C;admitdate->date [2000:Integer,1:Integer,1:Integer]) and the spatial location of the three digital postcode E1C is specified as pcode3(name->E1C;geometry->geo[EPSG4326,point[-64.8032256544,46.0988295816]]:Geometry).

5.1.Non-spaital semantic query

The non-spatial semantic query retrieves data sources based on non-spatial attributes, such as name, description, and time. With the ontologies and rules included in the OO jDREW engine, these kinds of queries can be accomplished by the top-down reasoning method. For example, a query is to find the related information of senior people with "Pneumonia_and_influenza" cases recorded by hospitals during the first two months of year 2000. This query requires the use of ontologies we described above, including the respiratory disease ontology and age ontology. From the respiratory disease ontology, the subcategory of "Pneumonia_and_influenza" cases should also be included in the query results. The age ontology defines the age range of seniors is above age 65. Therefore, we have defined the disease_query rule, which integrates the ontologies and other rules (e.g., earlier, later) to implement the query.

```
disease_query(disease-
        >?disease:Respiratory_diseases;agetype-
        >?agetype;startdate->?startdate;
        enddate->?enddate;  age->?age:Integer;
        gender->?gender; postcode->?postcode) :-
health_event(disease-
        >?disease:Respiratory_diseases;age-
        >?age:Integer;gender->?gender;
        postcode->?postcode; admitdate->?date),
age(agetype->?agetype; age->?age:Integer),
earlier(?date, ?enddate),
later(?date, ?startdate).
```

With the disease_query rule, the query results can be retrieved from the OO jDREW interface, as shown in Figure 4. This interface shows the related information of the patients that meet the query condition, and all the solutions can be iterated by clicking the button 'Next Solution'.

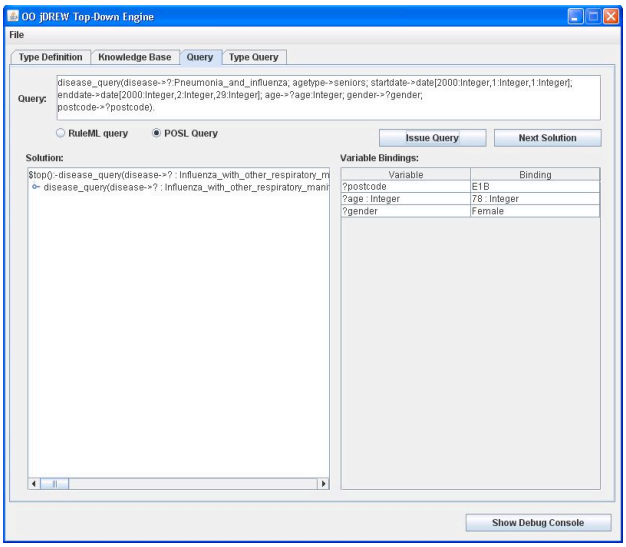

Figure 4. OO jDREW non-spatial semantic query result.

5.2.Non-spaital and geospatial semantic query

With the spatial information explicitly represented in the RuleML, spatial reasoning can be incorporated in health information retrieval. For instance, the above non-spatial query example can be restricted to find the patient cases within the geospatial boundary "Health region 1". The disease_locator rule is defined to support this query. The gpred_within built-in is used to determine whether the patient location is within "Health region 1".

```
disease_locator(healthregionname->?name;  disease-
        >?disease:Respiratory_diseases;  startdate-
        >?startdate;
        enddate->?enddate;agetype-
        >?agetype;age->?age:Integer;gender-
        >?gender;
        postcode->?postcode) :-
health_event(disease-
        >?disease:Respiratory_diseases;age-
        >?age:Integer;gender->?gender;
        postcode->?postcode; admitdate->?date),
age(agetype->?agetype; age->?age:Integer),
earlier(?date, ?enddate),
```

later(?date, ?startdate),

health_region(name->?name; geometry->?hrgeometry:Geometry!?),

pcode3(name->?postcode; geometry->?pcgeometry:Geometry!?),

gpred_within(?pcgeometry:Geometry, ?hrgeometry:Geometry).

Besides spatial relationship operators, the spatial operation operators can be used to generate new spatial features. For example, a query need to determine the correlation between smoking rate and income, such as finding the places with the smoking rate over 20% and average income value lower than 40000 in year 2003. A smoke_income_correlator rule is created to support this kind of queries, and the gfunc_intersection built-in is applied to calculate the geometry intersections of health region and census division that meet the non-spatial attribute condition. The result of this query is shown in Figure 5. New geometries have been generated for this query.

```
smoke_income_correlator(minsmokerate->?minsmokerate:Real;gender->?gender;agetype->?agetype;
                        year->?year:Integer;maxincome->?maxincome:Real;
                        geometry->?geometry:Geometry):-
  smoke_rate(geometryname->?dgeometryname;gender->?gender;agetype->?agetype;
             year->?year:Integer; rate->?rate:Real!?),
  income(geometryname->?igeometryname; year->?year:Integer;incomevalue->?incomevalue:Real!?),
  greaterThan(?rate:Real,?minsmokerate:Real),
  lessThan(?incomevalue:Real,?maxincome:Real),
  health_region(name->?dgeometryname;geometry->?hrgeometry:Geometry!?),
  census_division(name->?igeometryname;geometry->?cdgeometry:Geometry!?),
  gfunc_intersection(?geometry:Geometry,?hrgeometry:Geometry,?cdgeometry:Geometry).
```

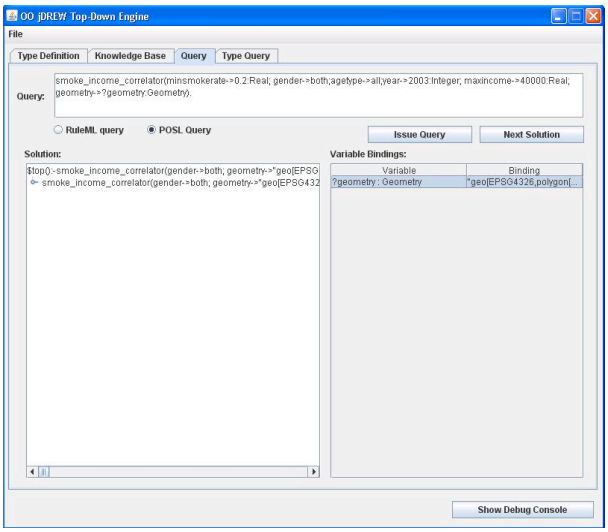

Figure 5. OO jDREW non-spatial and geospatial semantic query result.

Furthermore, we can include cartographic rules to determine how health information is to be represented in maps. For example, a specified color ramp is defined on how to show the smoking rate information on maps. The smoke rating is to be represented with three colors green, yellow, and red depending on its value with break point value of 0.2 and 0.5. Then, the smoke_rate_representation rule allows the integration this representation style to determine how to represent the corresponding results.

```
color_ramp(name->smokeramp;startvalue->0.000000:Real;endvalue->0.200000:Real;color->0x00FF00).
color_ramp(name->smokeramp;startvalue->0.200001:Real;endvalue->0.500000:Real;color->0xFFFF00).
color_ramp(name->smokeramp;startvalue->0.500001:Real;endvalue->1.000000:Real;color->0xFFFF00).
smoke_rate_representation(geometryname->?geometryname;geometry->?geometry:Geometry;
                          gender->?gender;agetype->?agetype;year->?year:Integer;
                          rampname->?rampname;color->?color):-
  smoke_rate(geometryname->?dgeometryname;gender->?gender;agetype->?agetype;
             year->?year:Integer; rate->?rate:Real!?),
  health_region(name->?geometryname; geometry->?geometry:Geometry!?),
  color_ramp(name->?rampname;startvalue->?startvalue:Real;endvalue->?endvalue:Real;color->?color),
  greaterThanOrEqual(?rate:Real,?startvalue:Real),
  lessThanOrEqual(?rate:Real,?endvalue:Real).
```

6. CONCLUSIONS

In this research the focus is on the non-spatial semantics of health data, using ontologies and rules. The geospatial component in the health data is incorporated in this study, and a geospatial-enabled approach has been proposed for semantic health information retrieval. The research proposes an architecture that applies ontologies, facts, and rules in health information reasoning and deduction from both geospatial and non-spatial dimensions. Ontologies and rules have been explored for the basic representation of health data from various sources in the Semantic Web. Spatial relation and operation operators are also enabled in the OO jDREW engine for spatial reasoning and knowledge discovery. This ontology and rule based health information integration and retrieval architecture provides initial exploration on how to utilize both non-spatial and geospatial semantics for health information retrieval and the case studies has demonstrated how the semantic query system works. Our future work will be on the enrichment of human knowledge as ontologies and rules for health data reasoning and deduction to make semantic query systems ready for real health applications.