Paulo César Barbosa Fernandes

An ontology-based system to generate epidemiologic profiles

Vitória - ES, Brazil

August 2012

Paulo César Barbosa Fernandes

An ontology-based system to generate epidemiologic profiles

Dissertation presented to obtain the Degree of Master of Computer Science of the Federal University of Espirito Santo.

Orientador:
Renata S. S. Quizzardi

Co-orientador:
Elias de Oliveira

FEDERAL UNIVERSITY OF ESPÍRITO SANTO
COMPUTER SCIENCE DEPARTMENT
MASTER IN INFORMATICS

Vitória - ES, Brazil

August 2012

Dissertation for Graduation Final Project under the title *"An ontology-based system to generate epidemiologic profiles"*, advocated by Paulo César Barbosa Fernandes and approved in August 2012, in Vitória, State of Espirito Santo, by the examination board consisting of the professors:

Renata Silva Souza Quizzardi, PhD
Universidade Federal do Espirito Santo
(Orientador)

Elias de Oliveira, PhD
Universidade Federal do Espirito Santo
(Co-orientador)

Giancarlo Quizzardi, PhD
Universidade Federal do Espirito Santo
(membro interno)

Fernando Silva Parreiras, PhD
Fundação Mineira de Educação e Cultura
(FUMEC)
(membro externo)

Abstract

Epidemiology is a field of study in Medicine which seeks to understand the factors that determine the frequency and distribution of diseases in humans. This field allows one to understand the phenomena of health and disease of a particular population by generating this population's epidemiological profile. The knowledge provided in this profile allows a shift in the focus from treating to preventing diseases, which is an important aim of the current Brazilian health care program.

This dissertation proposes a system to study the epidemiological profile in a basic health care unit. This system applies an ontology as basis for modeling and querying the epidemiological information. An ontology is a conceptual model which captures an specific view of a domain of discourse. This model may be used to structure the system's information, which later can be queried also with basis on this same ontology. A preliminary validation of this system's prototype has shown that it is able to successfully generate the health care unit's epidemiological profile, providing new knowledge about the patients and treatments involved in this unit. Such prototype may now be applied in this real setting to guide the actions of health care professionals in dealing with hypertension and other health conditions.

In order to develop the system, a goal-oriented methodology based on Tropos is applied. This methodology guides software development since an early stage of organizational modeling until the system's implementation by using current standards for ontology implementation.

Many of the available ontology engineering methodologies presuppose the existence of a set of questions which provide the objective and scope of the ontology under development. However, these so-called competency questions are not always clear from start. The highlight of the proposed methodology is applying goal analysis to assist the ontology engineer to reason about and model competency questions. Following this view, such competency questions are comparable to system requirements, elicited and modeled during the requirements engineering stage of a software development process.

Both the developed system and the proposed methodology are contributions of this work. However, while the former has proven to be useful in practice, further steps must be carried out in order to properly validate the latter, by applying it to other cases.

Keywords: competency question, goal modeling, i*/Tropos, ontology, C4.5, decision tree.

Dedication

“At my parents Paulo and Lizete
and my wife Leticia,
who helped me at all moments.“

Acknowledgements

“The Renata Guizzardi,
by example and attitude as guiding and teacher,
for his patience and guidance in performing this work.“

“The Elias de Oliveira,
by the teachings transmitted during the course of
Information Retrieval and Guided Study,
who contributed to this work.“

Table of Contents

List of figures

List of tables

1 *Introduction*

"I think if you show people
the problems and solutions,
motivate people to act."
Bill Gates

This chapter presents the purpose of this dissertation and the structure thereof.

1.1 Presentation

The Family Health Program (FRANCO; MERHY, 1999), was designed in 1994 with the goal to reorganize the health care practice with new bases and criteria. This program replaced the traditional care model, mainly aimed at curing diseases, in the hospital. With an emphasis on primary health care, this program organizes health services in order to be resolute, to minimize the economic and social costs, and especially to solve the population's health problems (CRUZ, 2007).

This program inaugurated the focus on the family as a unit of health programmatic action, and not the individual, introducing the notion of family doctor (VIANA; POZ, 2005), with the objective of health promotion. Currently, the Family Health Program is defined as the Family Health Strategy (FHS) instead of program, since the term program points to an activity with a beginning, development and completion and Family Health Program does not provide a time to complete this reorganization.

Epidemiology is a field of study in Medicine which seeks to understand the factors that determine the frequency and distribution of diseases in humans. This field has been important in enabling the FHS providing knowledge of the phenomena of health and disease. Such knowledge enables more proactiveness in health treatments. In other words, this kind of statistical information allows a shift in the focus from treating to preventing diseases, which is an important aim of the FHS.

Looking at the present computational support to knowledge modeling and discovering, the use of ontologies has become a popular solution in the past fifteen years. An ontology is a conceptual model which captures an specific view of a domain of discourse, for example, the domain of Epidemiology. This model may be used to structure information, which later can be queried also with basis on this same ontology.

This dissertation proposes a system to study the epidemiological profile in a basic health unit. As its name suggests, this system applies an ontology as basis for modeling and querying the epidemiological information. In order to develop such ontology, this work applies a methodology based on the use of Tropos to capture competency questions, which function as the requirements of the ontology-to-be. This methodology is also a contribution of the present dissertation.

1.2 Motivation

Characterized as an area strongly based on knowledge, medicine has been greatly helped by systems of knowledge management (KM) (STEFANELLI, 2004), dealing with the creation, integration and dissemination of knowledge, aiming to improve the performance of individuals and organizations. Among these systems, we highlight those aimed at diagnosis, health promotion and decision support, the planning of medical therapy and classification (BOSE, 2003). More recently, economic factors and quality also have become central for the medical KM.

Advances in information technology and telecommunications opened up new ways of dealing with knowledge, modifying the individual's relationship with work. The knowledge itself is seen as a product of great value, transforming the workplace in an environment of knowledge creation and continuous learning.

This has a big impact to the area of Computer Science, where professionals need to develop systems and methods that give support to the capture, dissemination, maintenance and exchange of knowledge.

Health institutions, in general, make use of various information systems, from those of nature management (eg patient registration system, payroll personnel, and others) to sophisticated knowledge-based systems that support the diagnosis, the monitoring of medical treatment and decision making in general. In addition, several non-automated processes are adopted to mediate tasks for which each member institution is responsible.

However, for knowledge to flow effectively in the organization, being led to the action points at the appropriate time, it is necessary to integrate the systems and processes. Thus, avoiding, for example, duplication of information, maintenance of old or false information, and especially the lack of information, which may be present in a system (or in the mind of one member of the institution) and absent in others, thus preventing that certain tasks are performed More effectively.

Given the need for an active vision for health intervention, not only focused on medical intervention (VIANA; POZ, 2005), epidemiology has become important to provide knowledge of the phenomena of health-disease.

Health-disease refers to all variables involved in health and disease of an individual or population. The concept of health-disease studies the biological, economic, social and cultural factors, seeking possible motivations for the appearance of illness, not considering as single factor of appearance of diseases only viruses, bacteria and protozoa.

Epidemiology is the science that studies the distribution and determinants of health problems in human populations (FILHO; ROUQUAYROL, 2002).

Currently, epidemiology is the main science of health information, revealing a strong vocation of science applied to the solution of health problems, consolidating scientific knowledge on human health, its determinants and consequences. In addition, many diseases, whose origins found no explanation until recently, have been researched by epidemiological studies (FILHO; ROUQUAYROL, 2002).

Given that the epidemiological profile is considered a relatively sensitive indicator of living conditions, the health-disease process and the development model of the population, the law 8.080 of 1990 (BARRETO, 2002) defines the Epidemiological Surveillance as "*a set of actions that provide knowledge, detection or prevention of any change in the determinants and constraints of individual or collective health with the purpose to recommend and adopt measures to prevent control of diseases and disorders*"(SUS, 1998).

At this point we emphasize the increasing importance of ontologies as a vocabulary for describing certain reality (GUIZZARDI, 2005), i.e. a model of a domain of discourse, as aforementioned. An ontology captures the concepts and relationships in a specific domain on a set of axioms that constrain the interpretation (GUARINO, 1998). Ontologies are used in various application domains to organize knowledge captured in order to facilitate access, understanding and allow the reuse of knowledge.

We also highlight the use of decision trees as a powerful tool for classification and prediction. The great advantage of decision trees is the fact that, in contrast to neural networks, decision trees represent rules. Rules can easily be expressed so that humans can understand them (QUINLAN, 1993).

In some applications, the accuracy of a classification or prediction is the only thing that matters. In such situations, it does not necessarily matter how or why the model works. In other situations, the ability to explain why a decision is crucial, as is the case of this work, where there is a need to explain the factors that led to the definition of a specific epidemiological profile. Through the use of decision tree, we can understand how the classification or prediction was found.

Thus, the motivation for this study is due to the importance of epidemiology for primary health care and the necessity to provide a strategy based on the proposal for the use of ontologies proposed by (GUARINO, 1998) and the use of decision tree (QUINLAN, 1993) to understand, first, how the work is performed in a basic health unit in the city of Vitória, through the proposed methodology based on Tropos (BRESCIANI et al., 2004) to analyze the competency questions that the ontology needs to respond, to then use ontologies together with decision tree in order to allow queries, deductions, inferences and development of rules that will assist in studying the epidemiology of the territory assisted by this given basic health care unit.

1.3 Research Objectives

The objective of this research is to propose a system that provides the epidemiological profile of a given region in Vitória (ES). This system is based on an ontology, which is also created in the context of this work. In order to create the ontology, we propose a methodology, based on the Tropos to analyze the competency questions that the ontology needs to respond.

1.4 Methodology

To achieve the objectives listed in the previous section, were performed the activities described below:

Firstly we performed a literature review related to ontology engineering, verifying how this issue is being addressed by the scientific community. In this survey, several studies have been developed to support the construction of these ontologies at this point we have identified limitations in these methodologies.

From this, the effort was to identify the best way to develop an ontology considering the use of the Tropos methodology. Using this language, we can model the actors of the organization, along with their goals and plans.

Furthermore, Tropos focuses on four phases of software development (early requirements, late requirements, architectural design and detailed design) and Tropos methodology is structured in terms of two main components: Actor Diagram and Goal Diagram. The former describes the organizational context in terms of dependency relationships between actors, while the latter describes the actors goals and rationales in order to justify the actors relationships.

After studying Tropos, we observed that the competency question (which are equivalent to the requirements if compared to requirements engineering) of the ontology can be identified after the stage of Late Requirements proposed in Tropos. So we proposed some extensions in the language, with the goal to use Tropos to derive competency question of the ontology.

Afterwards, we identifyied the concepts that the ontology should have, modeling it with the use of OntoUML (GUIZZARDI, 2005). The goal is to ensure that the language only accept as grammatically valid models those models that satisfy (the logical point of view) axiomatization of UFO (Unified Foundational Ontology), ie those models that are considered valid according to this theory. OntoUML is a modeling language that includes conceptual modeling primitives as the ontological distinctions proposed by UFO-A ontology.

To allow this ontology to be used to understand the epidemiological profile, through the use of an information system, it was fully built using Resource Description Framework (RDF) with the help of rules written in Euler[1] to make possible the execution of inferences and detection of inconsistencies in ontology queries using SPARQL, which is the query language for RDF.

Furthermore, with the goal of discovering new knowledge to enrich the ontology, we performed some experiments using decision tree, identifying rules to be added to the ontology. The insertion of these rules allowed new concepts to be treated, opening up new opportunities for consulting the ontology.

Finally, in possession of the ontology, with its defined rules, we developed an information system that uses the ontology in RDF model with the rules defined to allow SPARQL queries to be performed. Both RDF and SPARQL are standards related to the implementation of Ontologies. They have been widely used and proven useful in systems applying ontology based querying (QUILITZ; LESER, 2008)(BIZER; CYGANIAK; GAUß, 2007). Then tests have been developed for specific queries to support the activities performed by health professionals as well as to answer the competency questions identified during the development of the Tropos

[1]Euler Proof Mechanism - http://www.agfa.com/w3c/euler/

model.

Such prototype may now be applied in the basic health care unit focused in this work to guide the actions of health care professionals in dealing with hypertension and other health conditions.

As a result of our work, we were also able to generalize a methodology, composed of the steps that led us from the ontology conception to the implementation of the system. We also report on this methodology in this dissertation.

1.5 Organization of Dissertation

The remaining of this dissertation is organized in the following way:

- *Chapter 2 - Theoretical references:* presents a literature review of Primary Health Care, Epidemiological Profile, Tropos Methodology and Ontology, which were the basis for conducting this work.
- *Chapter 3 - Organizational Analysis:* presents the organizational analysis, performed in a case study in a basic health unit in the city of Vitória, using the Tropos language, as well as the adjustments made in the language to allow competency question to be captured.
- *Chapter 4 - An Ontology for Epidemiologic Profile of Primary Health Care :* presents the ontology created from the modeling goals, considering essencial concepts in the domain of basic health care.
- *Chapter 5 - A system for query on the epidemiological profile the Territory:* presents an information system that can perform queries on the ontology created, allowing information about the epidemiology of the territory to be consulted.
- *Chapter 6 - Final Considerations:* presents the conclusions of the work, contributions, limitations, and proposals for future work.

2 Theoretical references

"Science is what we understand well enough
to explain to a computer.
Art is everything else we do."
Donald Knuth

In this chapter we present the theoretical reference of concepts involved in this study.

2.1 Introduction

This chapter's main objective is to present the theoretical foundation for the work performed. Section 2.2 presents the Primary Health Care and the concepts that define it. Section 2.3 introduces the concept of epidemiological profile and its importance for Primary Health Care. Section 2.4 presents the Tropos methodology developed by BRESCIANI et al. 2004, which was extended in this work. Section 2.5 presents the concept of Ontology, its definition, the possible types of ontology, as well as the modeling language OntoUML (GUIZZARDI, 2005), in addition, a brief description of RDF which was the computer language used to build the ontology developed in this study. Finally, Section 2.8 presents the conclusions of the chapter.

2.2 Primary Health Care

The right of all Brazilian citizens to health was motivated, in 1980, by a broad national movement for health reform in the country, leading to a great change in the health care model in Brazil, the Constitution of 1988.

This section focuses on the primary health care, which is the context of this work. Subsequent sections describe a brief historical account that helps us understand how the current model of primary health care emerged, how the team of professionals responsible for it is composed and the main information system used as support.

2.2.1 Background

The public health was only included for, people who contribute to the National Institute for Medical Assistance of Social Security.

This institute had the responsibility to provide health care to its members only, which justified the construction of large units, ambulatory care and hospital, as well as the contracting of private services in large urban centers, which services were provided for workers of the formal economy (CARVALHO, 2001).

The increasing number of insured, and the need of increasing the supply of health services, lead to the construction of the network of hospitals linked to institutes of social security and medical assistance. Thus, private interests began to be evidenced in the state, producing an increase in medical and hospital services. Therefore, the medical practice with healing vision, intensified, became increasingly specialized, with elevation in drug use and their high cost (ANDRADE; SOARES; JUNIOR, 2001).

The occurrence of avoidable deaths, the persistence of infectious and parasitic diseases, the appearance of new diseases linked to lifestyle and industrialization, as well as threats of epidemics, expressed a low standard of health of the population. The public service, philanthropic and private, was shown to be unsatisfactory, leaving the population without having their health needs attended (PAIM, 2002).

Having this in mind, Brazilian health system was reformed in 1988, spurred significant changes in the management of the system. The reformed model of public health assistance occurred with the creation of the Unified Health System (*Sistema Único de Saúde - SUS*) (VIANA; POZ, 2005).

The new model, documented in the Constitution of 1988 established the principle of universalism for health actions, decentralization and a new organizational format for the services, under the logic of completeness, the regionalization and hierarchization, with the definition of the budget origins.

Moreover, the preventive and curative actions have become the responsibility of public managers (VIANA; POZ, 2005). Thus, it is considered that the organization of the Brazilian public health system is built mainly by the deployment of networks of *Basic Health Units*, in a more regionalized and hierarchical form, to conduct actions on the level of primary health care (BOTTI, 2006).

"Basic health care is characterized by a set of health actions, both individually and col-

lectively, which includes the promotion and protection of health, disease prevention, diagnosis, treatment, rehabilitation and maintenance of health" (MINISTÉRIO DA SAÚDE, 2009). Thus primary care is presented as a strategic component of the organization of health services, in order to be resolute, to minimize the economic and social costs, and especially to meet the health problems of the population (CRUZ, 2007).

The emphasis in primary care in Brazil means the possibility that the whole population is reached by this stage of care, and does not absolve from the Brazilian government to ensure people's access to all levels of care (CRUZ, 2007).

Then the Family Health Program (FHP) was created when the Department of Health created the Program of Community Health Agents (PCHA) in 1991. From there, the focus changed to the family as a unit of health program activity and not (merely) the individual, also introducing the notion of coverage area per family (VIANA; POZ, 2005).

The program also introduced a vision of active intervention in health, i.e. not expeting a demand to intervene, but to act preemptively about it, thus being an actual instrument for reorganizing the demand. In addition, another differentiating factor is the integration with the community and a focus not only centered on medical intervention (VIANA; POZ, 2005).

Nowadays, the FHP is defined as the Family Health Strategy (FHS) instead of program, since the term program points to an activity with a beginning, development and completion. The FHP is a strategy for the reorganization of the primary care and does not provide a time to finalize this reorganization.

2.2.2 Team Responsible for Primary Care

The FHS is formed by multidisciplinary teams that develops activities of promoting health, reception, home visits, health education, among others (CRUZ, 2007). Each team consists minimally of a family doctor (general clinical), a nurse, a nursing assistant and four to six Community Health Agents (BRAZIL, 2001). Each team is responsible for up to 4.000 people, being 3.000 the average recommended (BRASIL, 2006).

It's important to note, that the family health team, was inserted into the dental health care, through the Ministerial Decree no. 1.444/2000. The dental health team consists of dentists, technicians dental hygienists and dental assistants (CRUZ, 2007).

The Department of Health of Brazil (BRAZIL, 2001) requires each team to be able to:

- Know the reality of the families for which they are responsible, through registration and diagnosis of social, demographic and epidemiological characteristics.

- Identify the main health problems and risk situations to which the population it serves is exposed.

- Develop, with community participation, a local plan to address the determinants of the health / disease process.

- Provide integral care, responding to the continuous and rationalized, organized or spontaneous demand of unity in the community, at home and in monitoring to attend reference services in ambulatory or hospital.

- Develop educational activities to confront the health problems identified.

- Encourage the participation of the population widely, creating for this, discussion groups.

In this program, we highlight the *Community Health Agents*, which have the main specific tasks: mapping of residences, families registration and diagnosis, identification of risk areas, and realization of home visits giving priority attention to pregnant women and children, and community mobilization, in collective actions, and intersectoral action (CRUZ, 2007).

The *Community Health Agents* shall be responsible for up to 750 people, and there should be a sufficient number of agents to cover 100% of the population registered (BRASIL, 2006).

Another important task of the *Community Health Agents* is to maintain the health team informed about the families, especially about families in situations of risk.

The definition of risk factor that guides the work of the *Community Health Agents* is generally predetermined by health institutions, and reflects a strong influence of the ecological model, which refers to the links of the epidemiological chain, focusing on causation by a biological agent, or socioecological model, which shall include factors relating to behavioral and individual habits.

These data are collected by the *Community Health Agents* through, completion the forms for home care (BORNSTEIN; MARIA; LEAL, 2010).

2.2.3 Information System of Primary Care

To help monitoring and evaluating activities performed by *Community Health Agents*, aggregating and processing data coming from the home visits, as well as care, medical and nursing performed in basic health unit and at home visits, the Information System of Primary Care (Sistema de Informação da Atenção Básica - SIAB) was created in 1998 by the Department of Information and Informatics of the Unified Health System (DATASUS) (SILVA; LAPREGA, 2005).

This system represents a potential tool for the monitoring of registered families, as well as for local planning. However, the SIAB is still vertical and centralized, i.e., the flow follows the direction of the local to the central and the data analysis is still done primarily at the central level (Ministry of Health). Thus, the local and regional levels become only data transferer (SILVA; LAPREGA, 2005).

The problem mentioned above causes the SIAB to be rarely used by health teams as a tool for health planning, being more used only for survey about the number of some health conditions / disease (e.g. number of hypertensive, pregnant women, diabetics) for the performance of health groups in the community (SILVA; LAPREGA, 2005).

This work aims to improve health planning, allowing that data about the families, as well as the health conditions of its members, are analyzed into the basic health unit, allowing an epidemiological analysis to improve the planning of health teams.

2.3 Epidemiological Profile

Given the need for an active vision for health intervention and a focus not only centered on medical intervention (VIANA; POZ, 2005), introduced by the *Family Health Strategy*, epidemiology has become important to provide knowledge of the phenomena of health and disease[1].

Epidemiology is the science that studies the distribution and determinants of health problems in human populations. To this end, epidemiology is used for calculations and statistical techniques, not restricted to quantification, but also employing alternative techniques for the scientific study of collective health (FILHO; ROUQUAYROL, 2002).

The International Epidemiology Association (IEA[2]) in its "Guide to Teaching Methods", defines epidemiology as "the study of factors that determine the frequency and distribution of diseases in human communities. While the clinic is dedicated to the study of disease in the individual by analyzing a case, the epidemiology focuses on health problems in groups of

[1] refers to all biological, economic, social and cultural factors on the concept of health-disease studies and with them, is seeking possible motivations for the appearance of illness, not considering as single factor of appearance of diseases only viruses, bacteria and protozoa.

[2] IEA, http://www.ieaweb.org/, accessed on 07/04/2012.

people (...) most often involving large populations."

Nowadays the epidemiology is the main science of health information, and since its birth, has proved to be a strong vocation of science applied to the solution of health problems, consolidating scientific knowledge on human health, its determinants and their consequences (FILHO; ROUQUAYROL, 2002).

Many diseases, whose origins were not explained until recently, have been researched by epidemiological studies. We can cite, for example, leukemia in children caused by exposure to x-ray during pregnancy, infant mortality and social class; venous thrombosis related to use of contraceptives, among others (FILHO; ROUQUAYROL, 2002).

One of the great gains in this area was the discovery by John Snow that the risk of contracting cholera was related, among other factors, with the consumption of water supplied by a particular company. John Snow delimited the housing of each person who died of cholera in London between 1848-49 and 1853-54 and noted a clear association between the source of drinking water and deaths. From there, Snow made a statistical comparison of cholera deaths in each district and the source of obtaining water, showing that the number of deaths and the mortality rate were highest, among people whose water supply came from a certain company. Based on this research, Snow concluded that cholera was disseminated through contaminated water (BEAGLEHOLE; BONITA; KJELLSTROM, 2003).

Since the target of an epidemiologic study is human population, one must understand population not only as a set of individuals. In an epidemiological study, population may represent a specific group of hospitalized patients, or employees of an industry or people living in a geographic region among other possibilities.

In the vast field of public health, epidemiology is used in several ways. For BEAGLEHOLE; BONITA; KJELLSTROM (2003) the use of epidemiology can take four forms, these forms are presented in Figure 2.1 and detailed below.

- Causality - the causality of some diseases may be linked exclusively to genetic factors, such as phenylketonuria[3], but is more common as a result of an interaction between genetic and environmental factors. The behavior and lifestyle are of great importance in this connection, and epidemiology is increasingly being used to study its influence and the possibility of preventive intervention through health promotion.

[3]Phenylketonuria is a genetic disease, hereditary, and is characterized by the lack of an enzyme in greater or lesser proportions, preventing the body to metabolize and eliminate the amino acid phenylalanine. Consequently, too much blood is toxic, attacking mainly the brain and causing mental retardation and other problems.

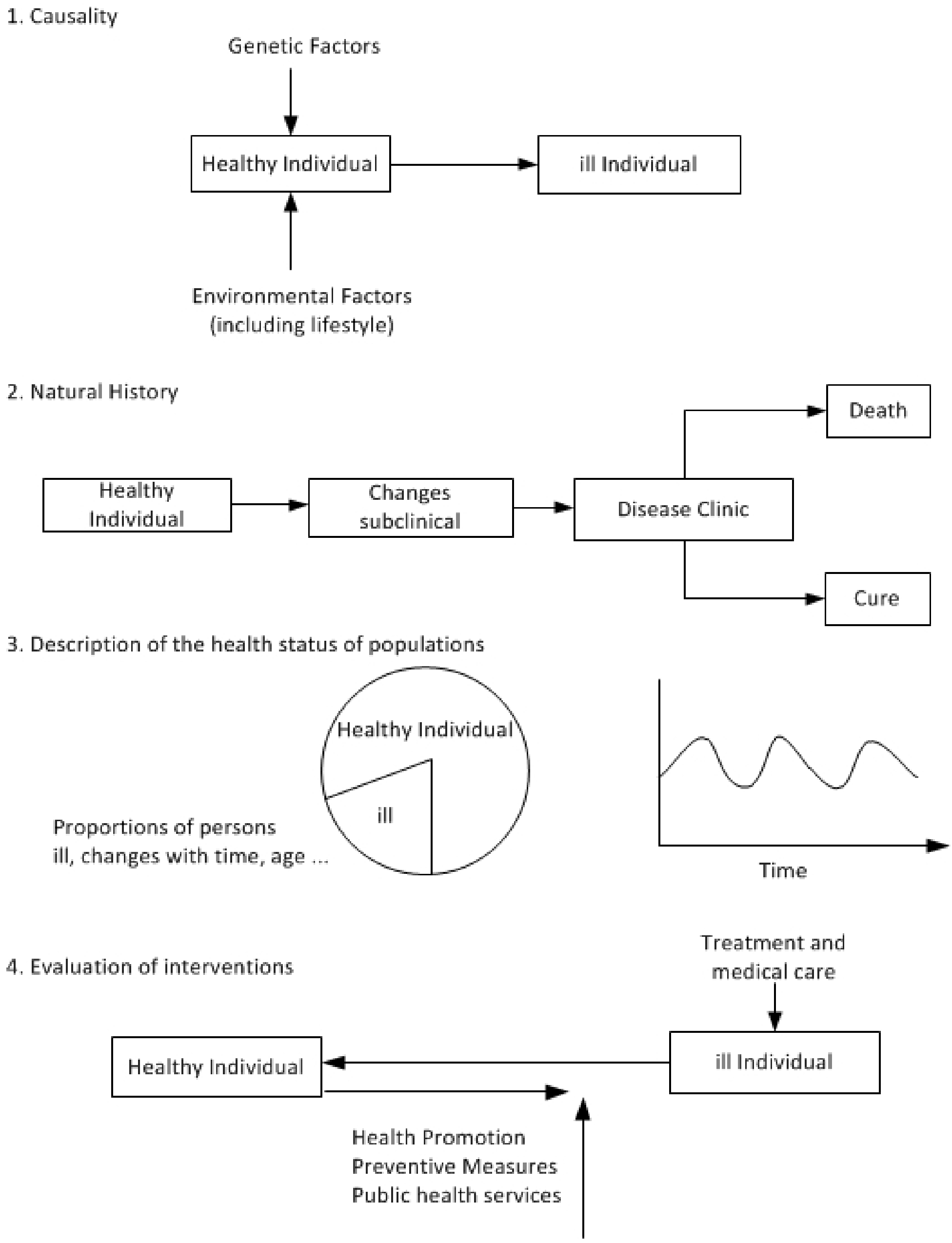


Figure 2.1: Uses of epidemiology. Source: BEAGLEHOLE; BONITA; KJELLSTROM (2003).

- Natural history - epidemiology is also preoccupied with the evolution and outcome of disease in individuals and population groups. Epidemiology, therefore, provides important insights for both curative and preventive medicine.

- Description of the health for populations - Knowledge of the disease that exists in the community is essential health authorities. Such knowledge allows limited resources to be used in the best possible way by identifying curative and preventive health programs, which are priority.

- Evaluation of interventions - the epidemiology involves the evaluation of the efficiency

> and effectiveness of health services. Examples of these services are to determine the appropriate duration of permanence in hospital for specific conditions, the value of treating systemic arterial hypertension, the effectiveness of sanitary measures in the control of diarrheal diseases and the impact on public health in reducing the levels of lead in fuel.

Epidemiology focuses on aspects of health that are relatively easy to measure and which are priorities for action. Definitions of health states used by epidemiologists tend to be extremely simple, such as "disease present"or "disease absent"(BEAGLEHOLE; BONITA; KJELLSTROM, 2003).

Thus, the criteria for determining the presence of a disease requires the development of normality and abnormality. Being the diagnostic criteria based on signs, symptoms and exam results, it becomes a difficult task to define what is normal or abnormal. Because the diagnosis can be made based on various manifestations of the disease, and some signals more important than others, there is no such clear distinction to determine the presence of a disease (BEAGLEHOLE; BONITA; KJELLSTROM, 2003).

Information systems with potential for use epidemiological information, that can or could be used in a decentralized level, has been a target of great discussion related to their conception. Until recently, all the epidemiological information systems were consolidated at the national and state levels. The cities only had access to their data after consolidation, that due to lack of agility in the availability of information, basic health units have great difficulty in working with these data (JUNIOR, 2003).

In some cases, the information systems focus on specific conditions, instead of a system that would comprise all diseases. Examples of such information systems are: SisHiperdia (for hypertension and diabetes), SISPRENATAL (for monitoring of pregnant women), SI-PNI (for information on vaccination / immunization), SISCAM (Information for Cancer in Women), SIAB (for information about Primary Health Care).

These systems, in its logic of conception, have a centralized view of the health system, with macros, delayed diagnosis or restricting the control instruments and especially the financial evaluation, which has a completely unsatisfactory role in contributing to health promotion and to the health professionals in a basic health unit.

2.4 Goal Modeling

2.4.1 Background

Goal-oriented approaches for Requirements Engineering (RE) have received increasing attention in recent years (KAVAKLI; LOUCOPOULOS, 2003). The adoption of these approaches has been motivated by the need of overcoming the semantic gap between a software system and the organizational environment in which the system functions.

In an attempt to reproduce the benefits obtained from applying goal-oriented approaches to Requirements Engineering, several goal-oriented approaches for business process modeling have been proposed recently (NEIGER; CHURILOV, 2004) (YAMAMOTO et al., 2006).

These approaches aim at reducing the semantic gap between the operational business processes and the organizational strategy behind these processes. To properly capture goals and strategies, these approaches commonly borrow methodologies and tools from the RE field.

This work uses the Tropos methodology to understand the actors, goals and their relations in a basic unit of health with family health strategy, with the goal of understanding the competency questions for development of an ontology.

2.4.2 Tropos Methodology

The i* modeling framework, firstly proposed in Yu's PhD thesis (YU, 1995), is an agent-oriented conceptual framework that can be used for requirements engineering, business process modeling, organizational impact analysis and software process modeling.

The i* framework has been used as a basis for the creation of the Tropos methodology and modelling language (CASTRO; KOLP; MYLOPOULOS, 2002)(BRESCIANI et al., 2004) which in turn, inherits the i* concepts in its early and late requirements stages and applies these concepts in the scope of an agent-oriented software development methodology.

The differences of Tropos and i* is that Tropos has been developed with focus on an agent-oriented paradigm for software development in mind (BRESCIANI et al., 2004), whereas i* has a broader applicability. Therefore, while Tropos focuses on four phases of software development (early requirements, late requirements, architectural design and detailed design), i* solely covers the two initial phases (BRESCIANI et al., 2004).

The Tropos methodology is structured in terms of two main components: Actor Diagram and Goal Diagram. The former describes the organizational context in terms of dependency relationships between actors, while the latter describes the actors goals and rationales in order to justify the actors relationships.

2.4.3 Actor

Actor is the agent-oriented concept representing an autonomous and social entity. An actor can represent a physical person, an organization, an organizational role, a type of intentional entity, a software system or component, or any other entity which has strategic goals and intentions, for example, an ontology based system. The framework adopts an intentional and strategic view of the organizational setting, relying on two basic assumptions (YU, 1995):

- Actors are intentional entities which accomplish tasks and activities based on motivations and rationales behind their practices;
- Actors are strategic entities which are concerned with the configuration of relationships established with other actors, adopting a particular configuration of relations;

The term actor is used for generically designating entities which can participate in intentional dependencies. To model sub-units of actors, there are three types of actors: agents, roles and positions.

- **agent** is an actor which displays a physical existence, such as human individuals or software agents, their characteristics are not easily transferable to other individuals.
- **role** is a characterization of a behavior of some actor in a given social domain, it is transferable to other individuals.
- **position** comprises in a set of roles which is performed by one agent. We say that an agent occupies a position and a position is said to cover a role (BRESCIANI et al., 2004).

2.4.4 Actor Diagram

The Actor Diagram reflects a network of strategic dependencies established between organizational actors, recognizing that actors have freedom of action. This model expresses the intentional perspective of business models instead of capturing the non-strategic concerns.

Each node represents an actor and each link represents the dependency between two actors. Dependencies can be characterized as follows: the depender is an actor which depends on some other actor (the dependee), by an intentional element (the dependum).

On one hand, by establishing dependencies, the depender is able to realize some goal that either it would not be able to accomplish or it would not be able to accomplish optimally. On the other hand, the depender becomes vulnerable in the sense that the dependee may fail to deliver the dependum.

The model has different types of dependencies based on how agents constrain each other's freedoms and how vulnerable the agents are according to the type of dependency established. There are four types of dependencies, based on the type of the dependum: **goal dependency**, **softgoal dependency**, **plan dependency** and **resource dependency**.

- **goal dependency** - A goal is a state of affairs desired by some agent. Hardgoals are defined as goals whose satisfaction can be determined by applying formal verification techniques (LAMSWEERDE, 2001) (JURETA; FAULKNER; SCHOBBENS, 2006).

  In goal dependency, the depender depends on the dependee to cause a desired state in the world, without prescribing the way in which the dependee must accomplish this state. The depender is vulnerable in the sense that the dependee may fail not satisfy this particular state. An example is shown in Figure 2.2.

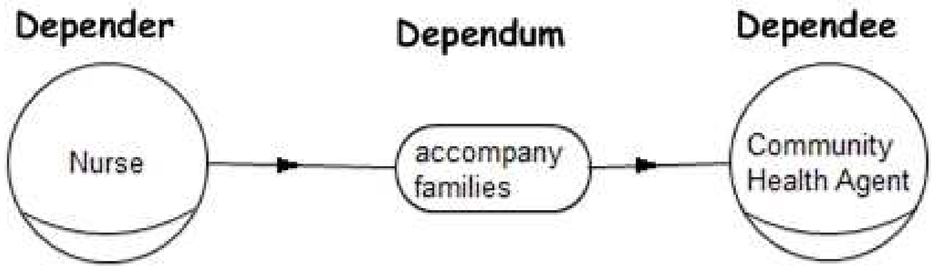


Figure 2.2: Example of (hard) Goal dependency.

- **softgoal dependency** - Softgoals are goals with no clear-cut criteria for satisfaction and thus, rarely, one can state that a softgoal has been satisfied. Instead, to refer to the satisfaction of a softgoal, one must adopt the notion of partial satisfaction, i.e., the softgoal is satisfied within acceptable limits. In this case, it is said that the softgoal has been "satisficed"(MYLOPOULOS; CHUNG; NIXON, 1992).Softgoals are the opposite of hardgoals, since they are "subject to interpretation"(YU, 1995), "imprecise, subjective"(JURETA; FAULKNER; SCHOBBENS, 2006).

  In softgoal dependency, the depender depends on the dependee to meet a softgoal. The softgoal achievement is settled during the trajectory of decomposing the hardgoal/task. The dependency also leads the depender to become vulnerable because the dependee can fail to bring about the softgoal. An example is shown in Figure 2.3.

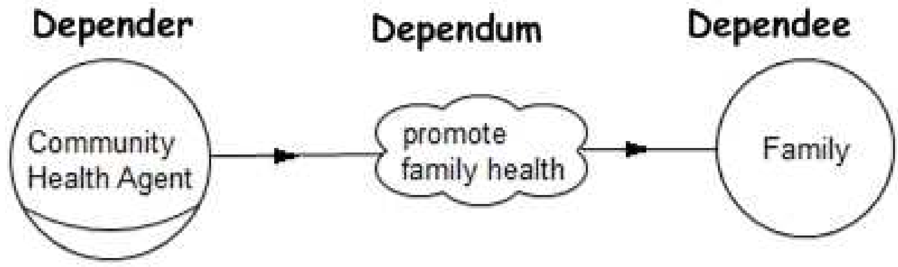


Figure 2.3: Example of Softgoal dependency.

- **plan dependency** - A plan represents a specific way of performing something, generally for achieving a goal or softgoal. In plan dependency, the depender depends on the dependee, again to accomplish a goal, but, this time, specifying the way in which the dependee must accomplish such goal. Thus, the dependee is constrained in its freedom of deciding how to accomplish the goal, i.e., the agent must adopt the specified plan to achieve the goal. The Figure 2.4 shows an example of plan dependency.

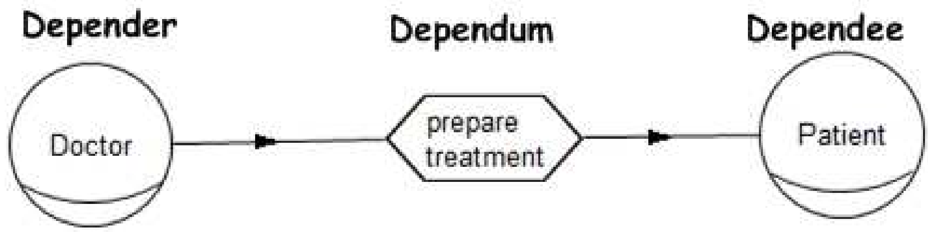


Figure 2.4: Example of Plan dependency.

- **resource dependency** - In resource dependency, the depender needs the dependee to deliver either a physical or informational resource, becoming vulnerable in the case of unavailability of the resource. Resources in the context of the framework must be seen as intentional objects (usually obtained as a finished product from a deliberation process), differing from flow of information which does not depicts intentionality. An example of resource dependency is shown in Figure 2.5.

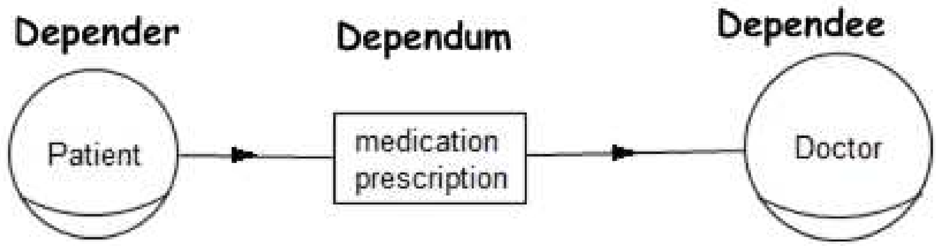


Figure 2.5: Example of Resource dependency.

2.4.5 The Goal Diagram

The Goal Diagram provides an intentional description of the process in terms of goals and relations that actors have in each business configuration. The description of these business configurations is elaborated in terms of intentional process elements. This model aims at expressing how actors reason when involved in a process of deliberation for adopting a particular configuration process. Moreover, process alternatives are explicitly expressed, allowing the systematic generation of different process designs based on the prioritization of concerns.

Goal diagrams have four types of nodes: *goals*, *softgoals*, *tasks* (or plan) and *resources*. And three types of relationships: *means-end* link, *contribution* link, and *AND/OR* decomposition.

- *Means-end* analysis aims at capturing plans, resources and softgoals that provide means for achieving a goal.
- *Contribution* analysis identifies goals that can contribute positively or negatively in the attainment of the goal to be analyzed.
- *AND* decomposition supports a goal to be decomposed in a series of sub-goals; while an *OR* decomposition allows modeling of alternative ways of achieving a goal.

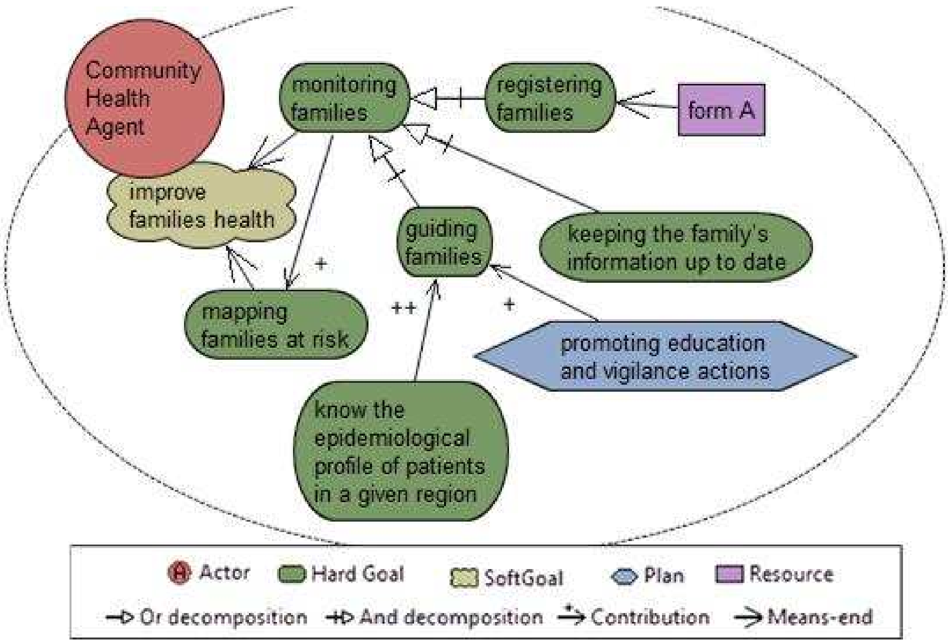


Figure 2.6: Examples of nodes and relationships of the Tropos methodology.

An example of Goal Diagram is shown in Figure 2.6 and described below. The ***Community Health Agent*** actor has a general softgoal of *"improve families health"*, which leads to the main goal of *"monitoring families health and life conditions"*. This goal is then decomposed in three important sub-goals: *"registering families"* at the **Family Health Strategy** if the patient is still not registered, *"keeping the family's information up to date"* and *"guiding families"* so that they can themselves act to promote and mantain their own health.

An important resource for accomplishing both the *"registering families"* and the *"keeping the family's information up to date"* goal is the *"form A"*, which contains all the relevant information about the family. This form is filled in in the first visit of a family member and should always be kept up to date. Updating the form will be necessary, for example, if some of the information (e.g. the address of the family, the number of family members, etc.) was changed since the last visit of a family member.

The *"promoting education and vigilance actions"* plan contributes to *"guiding families"*, as these families gradually learn to maintain pro-health habits and to play a more active role in keeping the authorities informed of any risk situation. The *"know the epidemiological profile of patients in a given region"* goal also greatly contributes to *"guiding families"*. This means that the ***Community Health Agent*** will understand the conditions of the region where the family lives, preventing disease before it proliferates among the other families in the region.

The *"monitoring families"* goal positively contributes to *"mapping families at risk"*. This is also a very important means for *"improve families health"*. For instance, after mapping the regions most affected by the dengue fever, the government could more intelligently direct resources to these regions. This could have a very fast positive effect on the number of cases in such regions.

This diagram reflects the goals, resources, softgoals and plans of each actor in each of the business processes executed by the department under consideration, in our example a basic health unit. Moreover, it also supports the capturing of general goals (and issues) which are not related with a specific business process. In each goal diagram, the strategic elements are properly related to show how change in one of them can affect the remaining elements.

Different versions of Tropos emerged, each using slightly different notations because of considerable attention from researchers and practitioners in recent years. The Tropos's models designed for this dissertation have been made with the support of TAOM4E[4] (GIORGINI et al., 2005). For this reason, we adopt the notation applied in this tool throughout this work.

[4]TAOM4E - available at http://selab.fbk.eu/taom/, accessed on 08/04/2012.

2.5 Ontology

The term *ontology* in the computer and information science literature appeared for the first time in 1967, in a work on the foundations of data modeling by S. H. Mealy, in a passage where he distinguishes three distinct realms in the field of data processing, namely: (i) "the real world itself"; (ii) "ideas about it existing in the minds of men"; (iii) "symbols on paper or some other storage medium"(GUIZZARDI, 2007).

Since the first time the term was used in AI by Hayes (HAYES, 1978) and since the development of his naïve physics ontology of liquids (HAYES, 1985), a large amount of domain ontologies have been developed in a multitude of subject areas. An explosion of works related to ontology has happened in computer science, chiefly motivated by the growing interest on the Semantic Web, and by the key role played by them in that initiative (GUIZZARDI, 2007).

2.5.1 Definition

For Gruber (1993), an ontology is a formal and explicit specification of a shared conceptualization. In this definition, conceptualization refers to an abstract model of a phenomenon in the world that identifies its relevant concepts. Explicit, because their concepts and restrictions on their use are explicitly defined. Formal because it is capable of being understood by machines. Shared reflects the fact that an ontology is able to capture consensual knowledge accepted by a community.

2.5.2 Types of Ontologies

Guarino (1998) suggests a classification into four types of ontologies:

- **top-level ontology** - describes very general concepts like space, time, object, event, action, and so on.
- **domain ontology** - describes the vocabulary related to a generic domain, eg, medicine, law, so.
- **task ontology** - describes the vocabulary related to a general task, for example, diagnosis, sales, so.
- **application ontology** - describes concepts dependent on a particular domain and task, which are often specializations of related ontologies.

The Figure 2.7 shows the relationship between the types of ontologies presented. Top-level ontologies are also known as foundational ontologies specialized. Furthermore, when these are combined, usually domain entities play roles while performing a certain task.

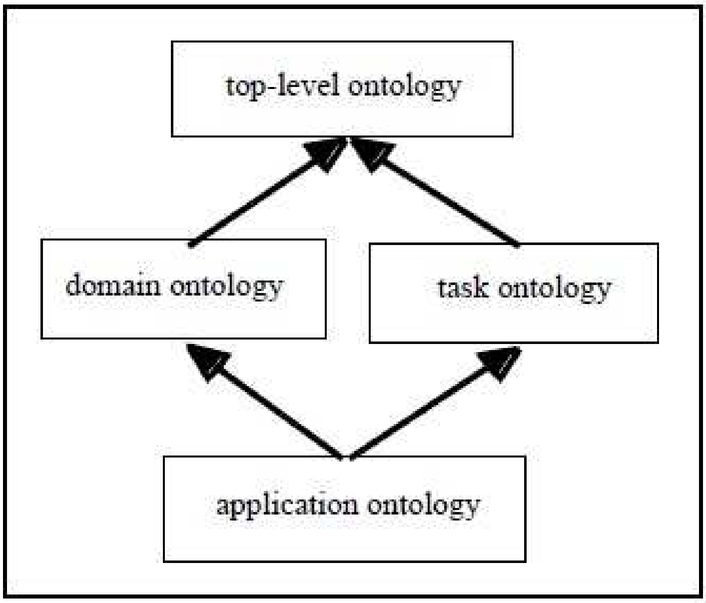


Figure 2.7: The four types of ontologies proposed by Guarino (1998).

Domain ontologies have been investigated extensively by the engineering community of artificial intelligence (AI), motivated by the need to enhance the reuse of software on a higher level of abstraction. A large amount of domain ontologies was developed for fields such as medicine, law, engineering, enterprise modeling, and chemical (GUIZZARDI, 2005). Also several methods for ontology engineering, focusing mainly on the domain ontologies have been proposed (GÓMES-PÉREZ; FERNÁNDES-LÓPEZ; CORCHO, 2004). To represent the domain ontology, in general, any structural model is used, such as UML class diagrams (FALBO, 2004) (CRANEFIELD; PURVIS, 1999).

2.5.3 Ontology Engineering

In Information Systems, the term ontology has been used in accordance with their definitions in philosophy, which is a system of categories independent of language. In contrast, in most other areas of computing the term ontology is generally used as an artifact of engineering designed for a specific purpose and represented in a particular language (GUIZZARDI, 2007).

In Guizzardi (2007) we are exposed to a series of questions: What exactly is a domain ontology? How does this relate to other concrete representations such as conceptual models and metamodels? How is it related to the ontology in the philosophical sense? Moreover, during the years many languages have been used to represent ontologies domain. What are the characteristics that a language must have to represent conceptual models, ontologies and domain? In particular, are the languages of the Semantic Web representation languages for ontologies appropriate?

To try to answer these questions, Guizzardi (2007) cites the need for two classes of representation languages in complementary Ontology Engineering: a language philosophically well founded, focused on expressiveness and conceptual clarity, and other languages with focus on computational concerns (for example, decidability, efficient automated reasoning, so on).

This idea is similar to that used in the discipline of software engineering, where there is a clear distinction between conceptual modeling, design and implementation.

- **In conceptual modeling** - an independent specification of a solution is produced, whose goal is to make a clear and precise description of domain elements for communication, learning and problem solving.
- **In the design phase** - this specification is transformed into a conceptual design specification, taking into consideration issues such as architectural styles, and non-functional criteria of quality to be maximized. The same conceptual specification can potentially be used to produce a number of different designs.
- **In the implementation phase** - a project is coded in a target language to then be implemented in a computing environment. Again, from the same design, different implementations may be produced.

At the stage of conceptual modeling in Ontology Engineering, highly expressive languages should be used to create ontologies that approximate the ideal of the domain ontology. The goal of these languages is the adequacy of representation, since the resulting specifications are intended to be used by humans in tasks such as communication, domain analysis and resolution of problems. The resulting domain ontologies, named *reference ontology*, must be used in an off-line basis to assist humans in tasks such as negotiation of meaning and establishing consensus. On the other hand, once users have already agreed on a common conceptualization, versions of a reference ontology can be created. These versions have been named in the literature as *lightweight ontologies* (GUIZZARDI, 2007).

Contrary to reference ontologies, lightweight ontologies are not focused on representation adequacy but are designed with the focus on guaranteeing desirable computational properties (GUIZZARDI, 2007).

We use the term ontology to describe what Guizzardi called reference ontology, we adopted the term as a domain ontology to describe lightweight ontologies.

In this study we adopt the following two-step process, conceptual modeling and design and implementation. Thus, we use the OntoUML language for conceptual modeling of domain ontologies, which is based in UFO (Unified Foundational Ontology). Finally, the implementation of the domain ontology was done using the RDF language (KLYNE; CARROLL, 2004).

2.5.4 Foundational Ontology and OntoUML

Foundational Ontologies serve as a basis for establishing consensus and negotiation among humans. They have been used successfully to improve the quality of modeling languages and conceptual models (GUIZZARDI, 2005).

A Foundational Ontology provides a vocabulary of concepts that are independent of a particular problem or domain. In this study, the Unified Foundational Ontology (UFO) (GUIZZARDI, 2005) was used as the basis for the creation of the reference ontology (domain ontology).

UFO is divided into three complementary parts: UFO-A is an ontology of enduring individuals (endurants) and is the core of the UFO, UFO-B is an ontology of events (perdurants) and finally, UFO-C is an ontology of social entities, built on parts A and B of UFO. Guizzardi (2005) proposed an ontologically correct UML profile call OntoUML. This profile is an extension of UML 2.0, ontologically well-founded and has a metamodel isomorphic to UFO-A. OntoUML can produce quality ontologies and use them as a conceptual model of high expressiveness.

To create the domain ontology used in this work, we applied the concepts defined in UFO-A. Below, we detail some of these concepts of UFO-A, based on (GUIZZARDI, 2005).

Some concepts have the characteristic of being necessarily applicable to their individual while they exist. This feature is defined in UFO-A as rigidity. For example, the concept Person is rigid, because all individuals classified as Person can not cease to be a person while exist.

If an individual person John, for example, stop being a person, he must necessarily cease to exist. However, the concept Student, is anti-rigid, since all individuals classified by this concept, eg John, may cease to be students and still continue to exist.

In OntoUML, a class with the **«kind»** stereotype is a concept with the property of rigidity and the characteristic of providing a principle of identity for their individual that allows them to distinguish and count them. A useful heuristic to reach such a conclusion is to check whether it is possible to count individuals. We may use, for example, the concepts of Car or Person, among others.

Considering fingerprint as the principle of identity provided to every instance of concept Person, this is the type **«kind»**, while the concepts Man and Woman are the type **«subkind»** because, besides being rigid, specialize the concept Person and thus it inherits the principle of identity.

On the other hand, a class with the **«role»** stereotype is a concept with the property of anti-rigidity, more precisely a concept that has participation in an event or a particular relationship. This is, something that can not always be assumed as true. For example, Student is a role for Person (Kind) by the existence of the relationship of enrollment to an instance of the type (kind) School.

Further, the **«phase»** stereotype is used to represent anti-rigid entities that constitute possible stages in the history. For example, child and teenager is a phase-partition (or state) of a Person.

The stereoptype **«collective»** represents instances that are collections. Examples include a deck of cards, a forest, a group of people, a pile of bricks.

Finally, we have the **«relator»** stereotype. Relators are individuals with the power of connecting entities; a flight connection, for example, founds a relator that connects airports, an enrollment is a relator that connects a student with an educational institution.

2.6 Resource Description Framework - RDF

At the present stage of the Web's evolution, most traffic on the Internet is between human consumers using Web browsers and content providers using Web servers. This means that to find and retrieve data, human intervention is often required. As businesses move more of their daily operations online, computer-to-computer peer services are growing (CANDAN; LIU; SUVARNA, 2001). Therefore the need is increasing for developing a model which will bring structures to descriptions of the Web content, thus creating an environment where the tasks such as searching the Web could be automated.

Many proposals were made to the World Wide Web Consortium(W3C) for representation of Web-related metadata. The RDF was created in 1997, RDF has drawn influence from several different sources. The main influences have come from the Web standardization community itself in the form of HTML metadata, the library community, the structured document community in the form of SGML (Standard Generalized Markup Language) and more importantly XML, and also the knowledge representation community. Other areas of technology also contributed to the RDF design such as object-oriented programming and modeling languages, as well as databases.

2.6.1 RDF Model

The Resource Description Framework (RDF) is an XML-based language for describing information contained in a Web resource. A resource can be a Web page, an entire Web site, or any item on the Web that contains information in some form. RDF enables the encoding, exchange, and reuse of structured metadata. It allows for metadata interoperability through the design of mechanisms that support common conventions of semantics, syntax, and structure. RDF makes no assumption about a particular application domain, nor defines the semantics of any particular application domain. The definition of the mechanism is domain neutral, yet the mechanism is suitable for describing information about any domain (CANDAN; LIU; SUVARNA, 2001).

RDF is a syntax independent model for representing resources and their corresponding descriptions. It provides a model for describing Web resources; i.e., objects that are uniquely identifiable by uniform resource identifiers (URIs). The resources are described using property names, which express the relationships of values associated with resources. Values may be atomic or may be other resources, which in turn may have their own properties.

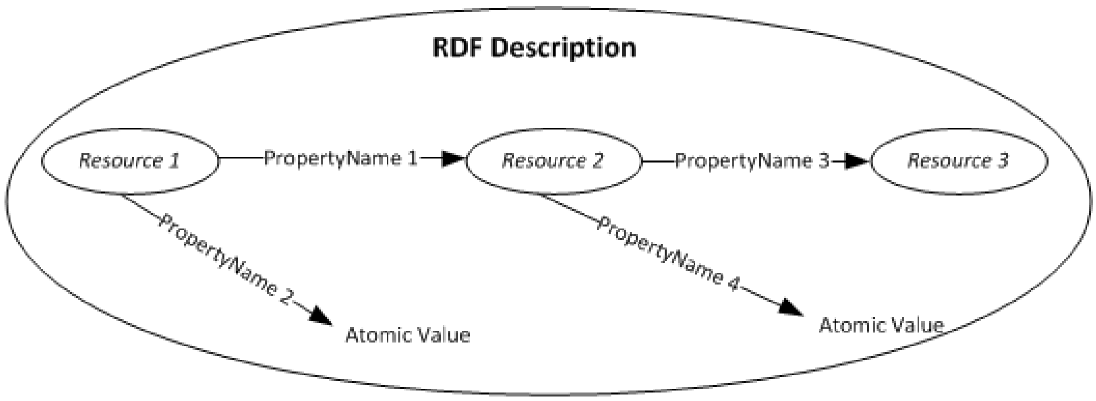


Figure 2.8: Overview of the RDF model.

A collection of these properties that refer to the same resource is called a description (Figure 2.8). Therefore, the RDF model consists of three major components:

- **Resources** - All things being described by RDF expressions are called resources. A resource may be an entire Web page, part of a Web page, an entire Web site, or an object that is not directly accessible via the Web page (e.g., a printed book).

- **Properties** - A property is a specific aspect, characteristic, attribute, or relation used to describe a resource. Each property has a specific meaning, defines its permitted values, the types of resources it can describe, and its relationship with other properties.

- **Statements** - A specific resource together with a property plus the value of that property for that resource is an RDF statement. These three individual parts of a statement are called the subject, predicate, and object, of the statement, respectively.

Let us consider the following statement about the patient John and see how we would use RDF to describe it.

Statement. "John has 30 years of age".

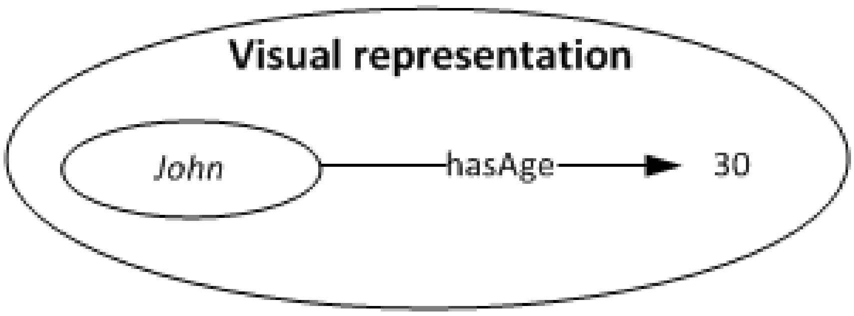


Figure 2.9: Visual representation of Statement.

Figure 2.9 shows how we can use RDF to express this statement using (1) a resource or subject (John), (2) a property name or predicate (hasAge), and (3) an atomic value or object (30).

2.7 Decision Tree

Decision trees are similar to if-then rules. They are commonly used in the implementation of expert systems and classification problems. The decision tree takes as input a situation described by a set of attributes and returns a decision, which is the predicted value for the input value. The input attributes can be discrete or continuous (RUSSEL; NORVIG, 2003).

The decision tree reaches its decision by performing a test sequence. Each internal node of the tree corresponds to a test of the value of properties, and the branches of the node are identified with the possible values of the test. Each tree leaf node specifies the return value if the leaf is reached.

To better understand the operation of a decision tree, let us consider the example in Figure 2.10. We consider the problem of correctly identifying the sex of a person from its name. The goal is to learn which sex from the person's name. In this example we disregard the trivial cases, such as, the sex of John who is always male.

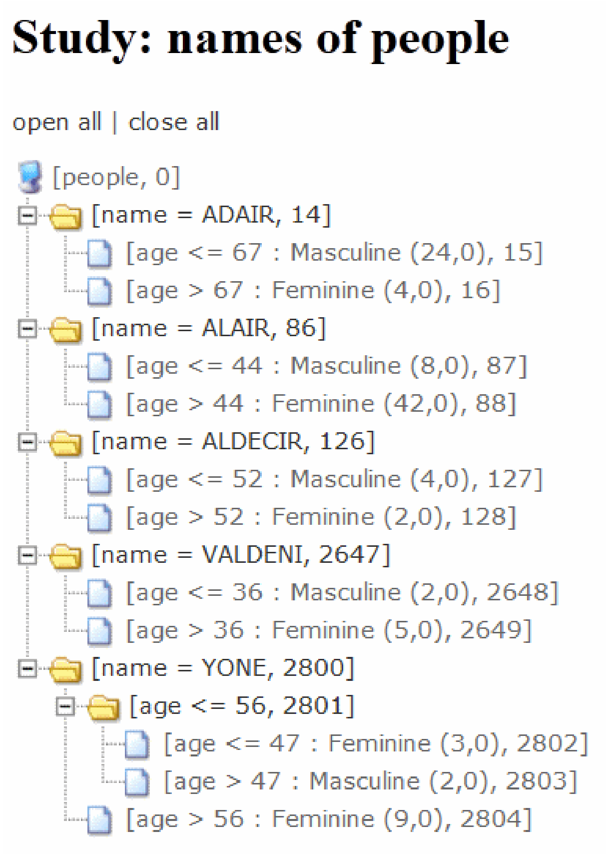


Figure 2.10: Example decision tree for the study of names of people.

For any problem of decision tree, one must first define attributes available to describe examples of possible cases of the domain. We take only the following attributes: name of the person, age and sex.

In a decision tree, the knowledge is represented in each node that, when tested, may lead the search to one of his sons. Thus, down from the root towards the leaves of the tree can be selected system configuration, and thus the behavior associated. In our example, if we wanted to search by name Alair, once found,should check the age, if greater than 44, the sex identified in accordance with the decision tree will be female (Feminine). This technique was used in step (e) the Ontology Engineering Approach proposed (see Figure 4.1), for creation of meta-rules.

2.8 Conclusions

This work aims to contribute to improving the health planning, allowing data about the families, as well as the health conditions of its members, to be analyzed within the basic health unit, providing an epidemiological analysis to improve the planning of health teams.

For this, we use the Tropos methodology to understand the actors, goals and their relations in a basic unit of health with family health strategy.

In this study we use the OntoUML language for conceptual modeling of domain ontologies. The ontology was encoded using the The Resource Description Framework (RDF) language (KLYNE; CARROLL, 2004). In this chapter, we presented a literature review, presenting the relevant works applied in this dissertation. We started by describing the Brazilian health system, then we presented the concept of epidemiology. Following, we described the Tropos language and, finally, we presented the concept of ontology.

In Chapter 3, we present the organizational analysis, performed in a case study in a basic health unit, using the Tropos language.

3 *Organizational Analysis*

"I value simplicity over everything;
I always look for simplicity."
Anders Hejlsberg

This chapter presented the organizational analysis performed in the Health Unit Santa Luiza.

3.1 Introduction

The rapid emergence of Knowledge Management as a key issue for improving the effectiveness of work practice (BROWN; DUGUID, 2000)(DAVENPORT; JARVENPAA; BEERS, 1996) should not be surprising. On the one hand, in fact, knowledge has always been recognized as an important component of any sense making practice; on the other hand, the growing complexity of work situations calls for more knowledge to avoid breakdowns.

In this work, the analysis of communities (professional of the unit Santa Luiza) and their main knowledge management requirements have been conducted through a combination of ethnographic[1] methods and action learning approach, where the focus is on the observation of the working practices and their analysis lead together with involved workers.

Using interactive approach with to activate user participation on system design as well as to build a mutual understanding among observers and observed workers.

Being the field analysis mainly focused on: the identification of the main knowledge exchanges among people, i.e., identification of knowledge network among experts in different business sectors (HOLLAND; LEINHARDT, 1979). Moreover, during the case study, representations of typical working scenarios, i.e., scenario-based design analysis (CARROL, 1995) have been used to support our work.

[1]Ethnographic - if consists of the study of an object by direct experience of reality where is inserted.

3.2 Case Study: Organizational Modeling the Health Unit Santa Luiza

This study was conducted in a basic health unit Dr. Jose Moyses (Health Unit Santa Luiza), located in the neighborhood Santa Luiza. The territory of operations the Health Unit Santa Luiza includes six neighborhoods of the city of Vitoria, they are: Barro Vermelho, Ilha do Frade, Itarare, Praia do Canto, Santa Lucia e Santa Luiza.

In Figure 3.1, are displayed the neighborhoods that compose the territory of operation the Health Unit Santa Luiza, separated into 25 micro areas[2].

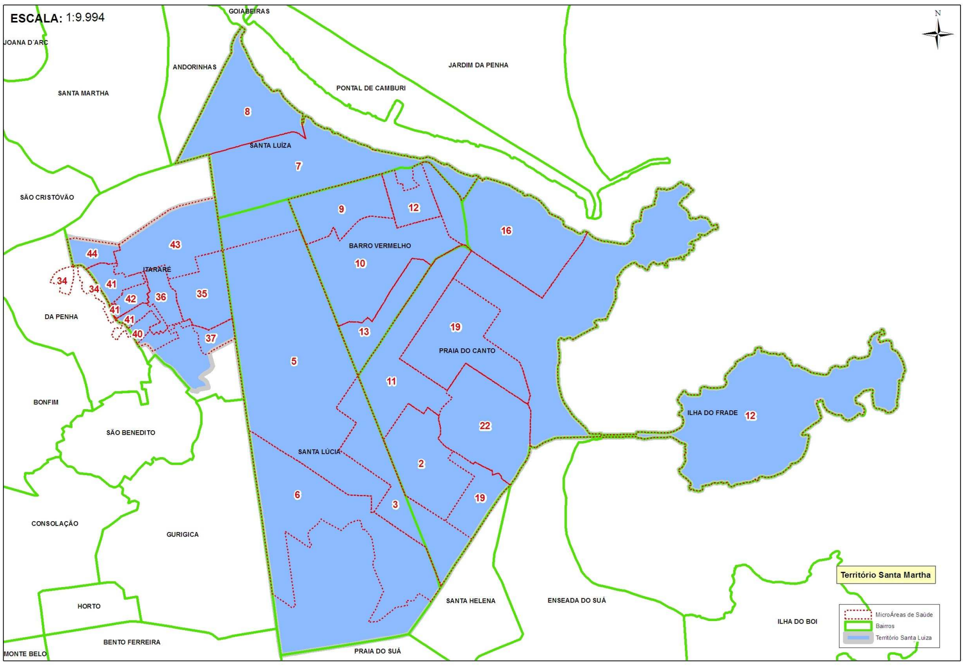


Figure 3.1: Health Unit Santa Luiza - Territory of health and its micro areas.

The neighborhood Itarare the largest number of micro areas, Figure 3.2, although not the most populous, this is due to the fact that the vast majority of the population is dependent on the public health system.

In Figure 3.2, we can observe that some micro-areas extrapolate the limits of the neighborhood Itarare, this is due to the need to better organize the number of families by micro area, in addition to improve the territorial organization, considering: streets, squares, among other features such as natural delimiters. Not being restricted only the geographic delimitation of the neighborhood.

[2]Micro Area - defined geographical space where they live about 400 to 750 people and corresponds to the operating area of a Community Health Agent.

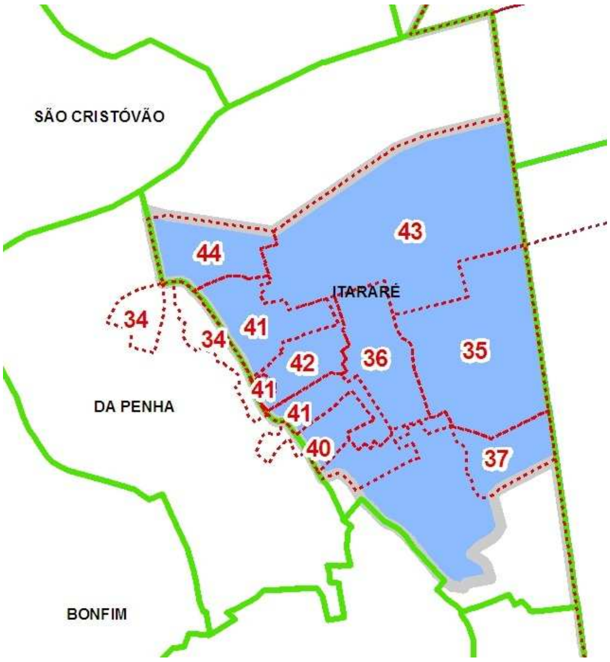


Figure 3.2: Micro areas that compose the neighborhood Itarare.

Table 3.1: Number of Families and People by health team.

Health Team	Number of Family	Number of People	Micro Areas
Team 1	1.146	3.561	2, 3, 7, 35 and 37
Team 2	1.321	4.052	9, 10, 16, 34 and 42
Team 3	1.334	4.129	6, 19, 36, 38 and 39
Team 4	1.285	3.821	6, 19, 36, 38 and 39
Team 5	1.402	4.392	5, 12, 13, 40 and 43
Total:	6.488	19.955	

To attend all this territory, making the monitoring of all registered families, the unit of Santa Luiza has 5 health teams, organized in accordance with Section 2.2.2. The number of families and people by health team as well as their micro areas, is shown in Table 3.1[3].

The organization of the health team, aims to facilitate patient access to health services by ensuring a better service allowing them to always consult with their professional, health team, it contributes to better monitoring. But, by no means the service will be denied to the patient even if the professional of his health team is unavailable for any reason, such as in the cases of: holidays and other cases.

[3]Data obtained by "Rede Bem Estar"system, health management system in the city of Vitoria.

3.3 Early Requirements: the Goals of the Community Health Agent and Nurse

In this section we will present the Tropos models created with the collaboration of professionals in the Santa Luiza unit.

These models were created from the analysis performed in the health unit. It has been observed how the concepts of the Family Health Strategy are put into practice. During this study, it was possible to identify very clearly the importance of Community Health Agents, as presented in section 2.2.2.

The Community Health Agents who live in the territory itself, are those who have primary responsibility, conducting home visits to monitor the families registered in the territory health. It is through these visits that the health team, gets to know how the health condition of families is.

Among some of the important observations made during visits and interviews conducted for this study, we highlight:

I Any problems related to any family of the territory is communicated to the Community Health Agent nurse in the team;

II The Community Health Agent aims to visit all families at least once a month;

III Since some Community Health Agent have a large number of families in its territory, it is not possible to visit every time each month. There should be a prioritization of visits per degree of risk families, families with a high degree of risk should be prioritized;

IV There are cases where some families are monitored almost daily, for example, when a family member had a surgery or when there is a need for special attention at the beginning of any treatment, and the responsibility of the Community Health Agent performing such monitoring under the supervision of the nurse;

V The doctors and nurses, among others, also do home visits, but only when there is such a request by the Community Health Agent, this usually occurs when, for example, a doctor should perform a visit at the patient's home, because the patient has some difficulty to move to the health unit;

VI An important activity performed by health professionals is to promote educational activities, such as for diabetic, hypertensive and smokers patients. These activities aim at explaining to patients about these diseases and how to treat them. These activities are possible with the help of the Community Health Agent, who knows the people and their health conditions.

VII The existence of protocols defined by the Department of Health, which sets the number of query to a pregnant woman during the period called pre-natal, also defines the hypertensive and diabetic patients should perform medical consultation monthly, and so on. In order to ensure that these protocols are followed, it is the responsibility of the Community Health Agent on each visit to check whether patients are appearing the in the unit to appointments.

Looking at the list of observations mentioned above, we notice the importance of the Community Health Agent for Family Health Strategy. Since the goal of the Family Health Strategy is not expected the problem to happen, but act with preventive health, the Community Health Agent has this important role of monitoring and accompanying the families under their responsibility.

But for this to occur, not only the Community Health Agent, but the entire health team must be well informed about the health conditions of the families under their responsibility, and it is not easy.

Looking at Table 3.1, considering the *Team 1* that has a smaller number of families under their responsibility, and divide the number of families by the amount of Community Health Agents, who are five (one for each micro area), we obtain the value of 229 families per agent. Whereas, 20 working days in the month, each agent has to make 11 visits per day.

But besides the visits the Community Health Agent, accompanies other professionals in educational activities, primarily accompanying families most in need with more than one visit in the month in addition to other factors that may influence the itens (II e III) are achieved.

If we also observed in the item (VI), it is not an easy task to perform an educational activity with the objective of health promotion. Let's consider an example of educational activity on hypertension. The health professional can select for this activity hypertensive patients that the Community Health Agent selected, either through reports made on paper or with the help of an information system, but the big question is, which patients are not yet hypertensive, but have characteristics (sex, age, place of residence) similar to those that were selected? These may be patients who should also participate so that they may not have in the near future this health condition.

After better understanding how the health unit of Santa Luiza works, we present below models created in Tropos. With the purpose of focus more on health promotion, the models are presented as main actors: the Community Health Agent, Doctor, Nurse and the Family.

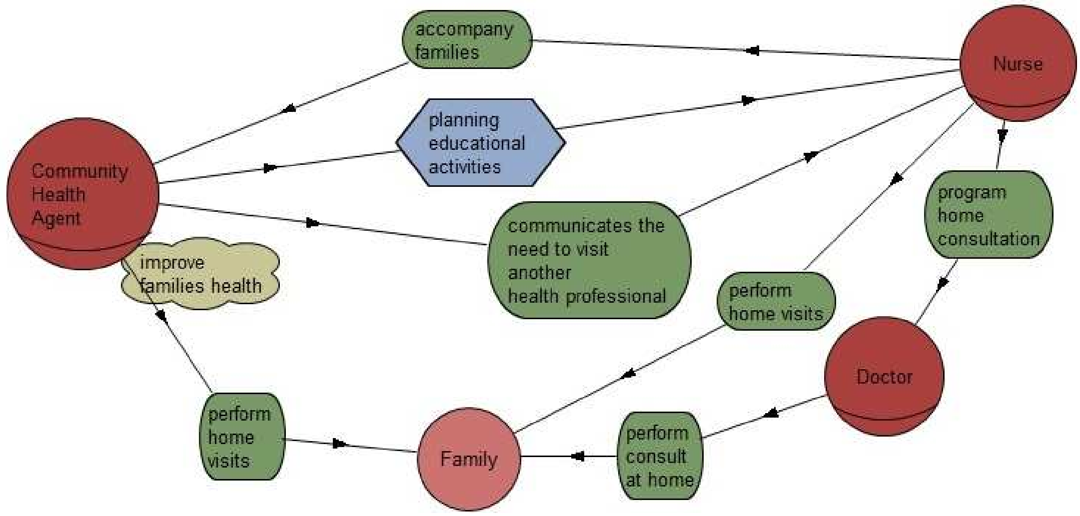


Figure 3.3: Early Requirements Unit Health Santa Luiza.

In Figure 3.3 we can view the main actors in the context of the health unit Santa Luiza and its dependencies. The actor **Nurse** who is responsible for **Community Health Agents** in her health team, depends on its **Community Health Agents** for all families, belonging to the micro areas the team, must be accompanied, represented by the goal *"accompany families"*.

The **Community Health Agent** that has as softgoal *"improve families health"*, has as main task to accompany families through home visits. Through visits to **Community Health Agent**, get to know families, and health condition of each of its members. Knowing well the health condition of patients under his responsibility the **Community Health Agent** can plan educational activities *"planning educational activities"* together with the **Nurse**, to promote the health of these families.

Another important objective of the **Community Health Agent** is *"communicates the need to visit another health professional"* whenever a family needs to visit another health professional, it is the responsibility of the **Community Health Agent** notify the **Nurse** on their team.

The actor **Nurse**, knowing the necessity of the visit of another health professional's family indicated by the **Community Health Agent**, he must evaluate the need for this visit, if there is a need for the **Doctor** to perform this visit the **Nurse** should *"program home consultation"* with the **Doctor**.

If it's a home visit that the **Nurse** can perform, there is no need to schedule the visit of a **Doctor**, in this model is shown by the dependence between the **Nurse** and **Family** through the objective *"perform home visits"*.

The actor in the model **Family**, does not represent a particular family, but any family of the Health Territory.

The actor depends **Family** health professionals whether they are: **Doctor**, **Nurse**, **Community Health Agent** or other, to be accompanied through home visits. This is an important point where the **Family** has a particularly strong dependence with the **Community Health Agent**, if this dependency fails, the model is notable that the entire proposal from the Family Health Strategy will fail, and the goals will not be achieved.

Knowing the main actors and the dependencies between them, the next models will highlight the goals and relationships of the actors **Community Health Agent** and **Nurse**, the choice of actors is because they are important parts of the Family Health Strategy. Show the objectives and tasks performed by these actors to the proposal from the Family Health Strategy is achieved.

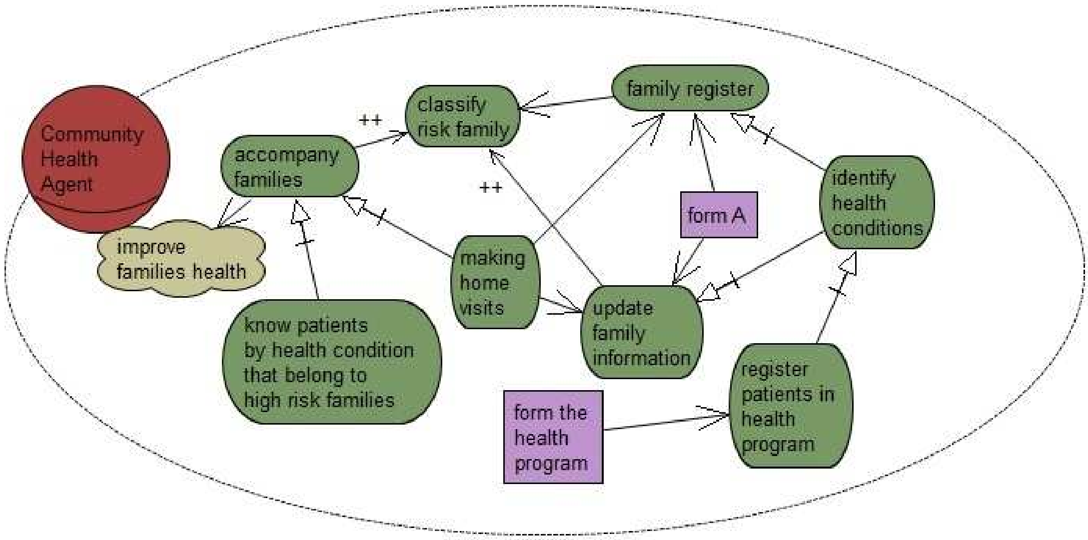


Figure 3.4: Goal diagram **Commnunity Health Agent**.

In Figure 3.4, is shown the Goals diagram of the actor **Community Health Agent**. The **Community Health Agent** has as main goal, *"accompany families"*, this goal is a means (***means-end link***) for the softgoal *"improve families health"* be achieved.

The objective *"accompany families"* is decomposed into two other goals *"know patients by health condition that belong to high risk families"* and *"making home visits"*. It is through the goal, *"making home visits"* the **Community Health Agent** *"family register"* and also *"update family information"* families have already registered.

Through the goal, *"family register*"the **Community Health Agent** must complete the registration form the family, represented by the resource, *"form A*". At this moment must also be identified the health conditions of the family members, (hardgoal *"identify health conditions"*), depending on the health condition of the patient it must be registered with a specific program (hardgoal *"register patients in health program"*), for example, for hypertension and diabetes there a program called Hiperdia, to deal exclusively with these health conditions.

Even if a family already was registered by **Community Health Agent**, during a home visit, he must make sure that there is a need for *"update family information"* do not have the correct information and updated the family will affect the monitoring of this family. For example, a pregnant woman to the incorrect date of birth can define whether her pregnancy is at risk or not.

For item II (The Community Health Agent aims to visit all families at least once a month) it is possible to reach the **Community Health Agent** shall classify the risk of their families so that families of high and medium risk are always accompanied.

Through the goal, *"classify risk family"*, the **Community Health Agent** shall classify the risks of family according to their health conditions and housing the basic health unit Santa Luiza, the scale used for this purpose is called Scale of Rabbit (COELHO; SAVASSI, 2004). This scale defines the risks of the family as: no risk, low risk, medium risk and high risk.

The goals, *"accompany families"* and *"update family information"* contribute significantly (***contribution link***) to the risk classification is correctly reflecting the health status of families. Accompanying families the **Community Health Agent** is updated on the health status of families in addition, you can have the family information always updated, contributing correctly to the risk classification of families.

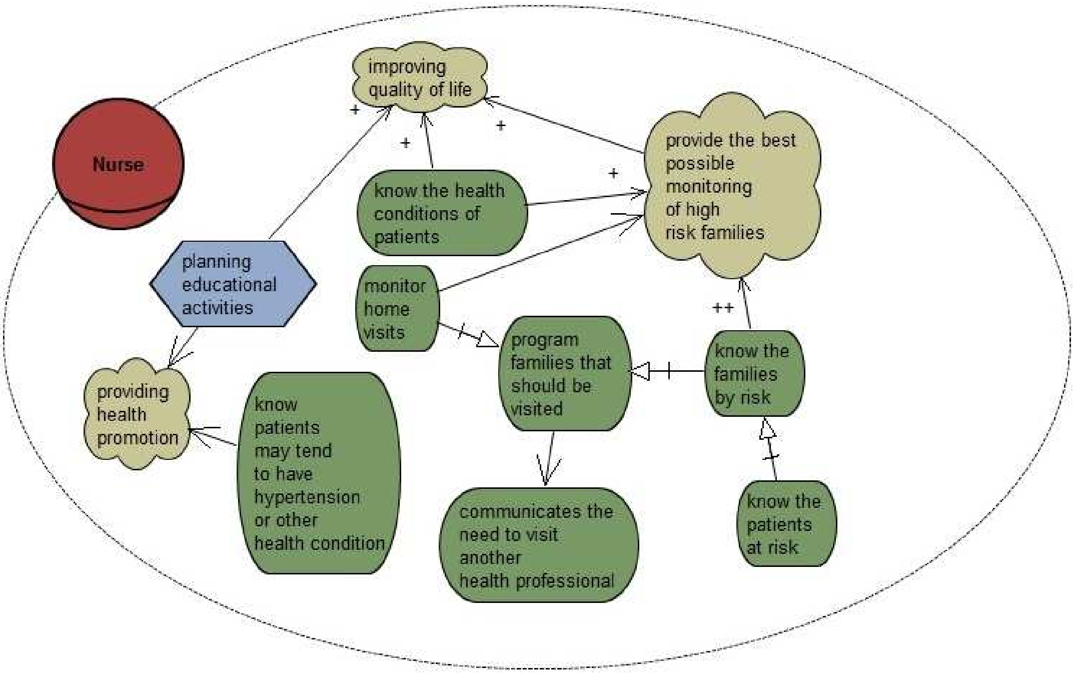


Figure 3.5: Goal diagram **Nurse**.

In Figure 3.5, is shows the goals diagram of the actor **Nurse**. The plan *"planning educational activities"* is an important task performed by the nurse. Through this educational activity, the nurse seeks to contribute to *"improving quality of life"* of patients participating in the educational activity, informing or guiding the participants.

The plan *"planning educational activities"* is a means to *"providing health promotion"*, but for the educational activities that have characteristics of health promotion is important to the effective participation of the **Community Health Agent**.

The **Community Health Agent**, during the visits can detect the incidence of a type of health problem, common among the families, for example, the incidence of diarrhea, which could be caused simply by consuming unfiltered water or not boiled.

An educational activity performed in the community could prevent such problems to continue to occur, but for this to occur it is very important to the perception by the **Community Health Agent**, identifying opportune moments to perform educational activities and informing the **Nurse** about it.

An important task that must be performed by the **Nurse** is *"program families that should be visited"*, by **Community Health Agent**. The goal, *"program families that should be visited"* is decomposed (***and decomposition link***) into two other goals: *"monitor home visits"* and *"know the families by risk"*.

The **Nurse** knowledge of families that have been monitored by the **Community Health Agent** and knowing families by degree of risk, can plan visits that must be performed by **Community Health Agent**, the next week or fortnight. Also can identify families who will need a visit from some other professional and *"communicates the need to visit another health professional"*, for example, a doctor or psychologist.

The goal, *"know the families by risk"* has a subgoal, *"know the patients at risk"*. It is very important for the **Nurse**, when perform the programming of the visit, besides to know the degree of risk families also know the patient for risk. At times a family with low degree of risk, may have a member who has some health condition, which is being investigated, then this patient should be monitored.

One goal, always pursued by the nurse is the softgoal *"provide the best possible monitoring of high risk families"*, the goal is to always accompany the high-risk families, seeking them to leave if possible to be high-risk families. In some cases this will not occur, which is the case of chronic diseases or incurable diseases, where the goal of accompanied, is to give quality of life for these families, ***contribution link*** with softgoal *"improving quality of life"*.

Another important and difficult goal to be achieved by **Nurse** is *"know patients may tend to have hypertension or other health condition"* in the city of Vitória, there is, to the role called Nurse Sanitarian, which seeks to understand how diseases are growing in population and discuss strategies to attack the problem. Would answer questions like, what are the characteristics (age, sex, so on) of people who are doing drugs in micro area 19? And what are people prone to start doing drugs in the same region?

Able to answer these questions is a way to promote health promotion (***means-end*** with softgoal *"providing health promotion"*), avoiding, for example, that drug use will destabilize a family, causing even a more serious problem.

3.4 Conclusions

In this chapter we present the study conducted in a basic health unit Dr. Jose Moyses (Health Unit Santa Luiza), located in the city of Vitoria. With the collaboration of professionals in the Santa Luiza unit, the Tropos models were created.

These models were created from the analysis performed in the health unit. Have been observed how the concepts of the Family Health Strategy are carried on. During this study, it was possible to identify very clearly the importance of Community Health Agents.

Since the goal of the Family Health Strategy is not expected the problem to happen, but to act with preventive health, the Community Health Agent has this important role of monitoring and accompanying the families under their responsibility.

The Community Health Agents who live in the territory itself, are those who have primary responsibility, conducting home visits to monitor the families registered in the health territory. It is through these visits that the health team, gets to know how are the health conditions of families, under their responsibility, and it is not easy; considering the number of families under the responsibility of Community Health Agent.

Furthermore, the objective of health promotion, which is performed through educational activity, is not an easy task. Let's consider an example of educational activity on hypertension. The health professional can select for this activity hypertensive patients that the Community Health Agent selected, either through reports made on paper or with the help of an information system, but the big question is, which patients are not yet hypertensive, but have characteristics (sex, age, place of residence) similar to those that were selected? These may be patients who should also participate so they do not have in the near future this health condition.

From the initial study presented in this chapter, the Chapter 4, presents the ontology created from the modeling goals, considering the concepts essences for basic health care. Assisting, for example, the health professional in the epidemiological study of the territory of health.

4 *An Ontology for Epidemiologic Profile of Primary Health Care*

"*When you are clear, what you want will show up in your life,*
and only to the extent you are clear."
Janet Attwood

In this chapter we present the organizational analysis, Late Requirements phase, as well as the ontology created using OntoUML.

4.1 Introduction

In recent years, we have been watching an explosion of works involving the use of Ontologies, mainly motivated by the need for semantic interoperability among domain models and applications; this is due to growing interest on the Semantic Web (GUIZZARDI et al., 2009). Practical benefits of an ontology-based system are, for example: a) allowing developers to reuse knowledge instead of reusing software; b) enabling the developer to reuse and share application domain knowledge using a common vocabulary across heterogeneous software platforms and programming languages; and c) letting developers focus on domain structure, preventing them from being distracted by implementation details.

To provide the aforementioned benefits, Ontologies are commonly used in computer science either as a reference model to support semantic interoperability, or as an artifact that should be efficiently represented to support tractable automated reasoning. We argue that in either case, a consistent ontology engineering methodology must be applied to develop Ontologies in a systematic way.

In this context, it is important to highlight that Ontologies inherently have a social nature, i.e. they are typically regarded as an explicit representation of a shared conceptualization, i.e., a concrete artifact representing a model of consensus within a community and a universe of discourse. That is why an ontology development begins with the definition of a set of questions, named competency questions, defining its objective, scope and expressiveness requirements. In other words, these are the question that a particular community of users thinks the ontology under development should answer.

In a sense, we can draw a parallel between an ontology's competency questions and the requirements of a system. If we think carefully, both a competency question and a requirement identify a future characteristic of the ontology (the former) and the system to be (the latter). In software, this is then translated into functionality while in an ontology, this is materialized by the right set of concepts and relations.

It is known that the requirements engineering activity must be thoroughly and precisely executed to guarantee the success of a software development project, thus being critical to assure software quality. Boehm (BOEHM, 1987) affirms that "Finding and fixing a software problem after delivery is 100 times more expensive than finding and fixing it during the requirements and early design activities."What Boehm means is that omissions, contradictions, ambiguity, not discovered during the initial requirements will propagate to the other software development activities. And this can later result in a system full of bugs or in the need for an expensive and time-consuming redesign.

If we go back to the comparison between competency questions and requirements, we can than conclude that the elicitation and modeling of competency questions should also undertake a thorough and precise process. Curiously, this is not the case so far. Most ontology engineering methodologies presuppose the existence of this set of questions and concentrate their efforts in defining the next stages of ontology development.

This work is an attempt to provide a solution to fill in the gap between the definition of the competency questions and the ontology modeling per se. For that, we rely on goal modeling.

Goal modeling has been successfully applied in requirements engineering in the past years (GUIZZARDI; PERINI, 2005)(LIASKOS et al., 2010) to deliver software for organizations in general. One of the main strengths of this approach comes from the importance it gives to the understanding of the current state of the organization (early requirements, presented in section 3.3) before a solution is actually conceived. The analysis of the goals of the actors composing such organization allows one to answer questions such as: "What are the strategic objectives of the organization?", "Who share these goals?", "What alternatives are considered for achieving the goals?"and "What are the plans and resources applied to achieve such goals?"

Instead of solely focusing on system requirements, our proposal encourages the use of goals, plans and resources to capture the competency questions that will contribute to achieving the organizational goals. From such competency questions, some initial concepts are drawn and, from then on, the ontology can go through iterative conceptual modeling stages.

In this chapter, we apply Tropos (BRESCIANI et al., 2004), in phase Late Requirements to enable the definition of competency questions.

4.2 Motivation: Why ontology-based systems

Ontologies provide a number of useful features and serve multiple purposes in intelligent systems, as well as in knowledge representation and knowledge engineering in general. Gasevic, Djuric and Devedzic (GASEVIC; DJURIC; DEVEDZIC, 2009) describe two of these features: vocabulary and taxonomy.

An ontology provides a vocabulary for presenting the terms of a given domain in the way these terms are understood and applied by a community. This vocabulary comprehends logical statements that describe what the terms are and how they can or cannot be related to each other. Additionally, an ontology aims at specifying terms with unambiguous meanings, i.e. having their semantics independent of a particular reader and context. Thus, to sum up, an ontology-based system provides a vocabulary and a machine-processable common understanding of the terms of a particular domain.

Another interesting feature is Taxonomy. Taxonomy is a hierarchical categorization or classification of entities within a domain. It may, for example, serve the purpose of clustering entities based on common ontological characteristics. An ontology is inherently composed of a taxonomy, which may be codified in a machine-readable and machine processable form.

The ontology's vocabulary and taxonomy together provide a conceptual framework for discussion, analysis, and information retrieval in a given domain (GASEVIC; DJURIC; DEVEDZIC, 2009)(GUIZZARDI; LUDERMIR; SONA, 2007). Bringing this to the health care domain, for example, they can support professionals and patients to acquire the right piece of knowledge in face of a certain health care problem or situation.

Gasevic, Djuric and Devedzic (GASEVIC; DJURIC; DEVEDZIC, 2009) also present knowledge sharing and knowledge reuse as two of the main purposes of a domain ontology. Both sharing and reuse are possible because of the common vocabulary provided by the ontology. Even if you do not consider any automation, having a common understanding of domain semantics enables people to communicate and cooperate better (GUIZZARDI et al., 2009). Moreover, software applications may also benefit from machine processable ontological models, which allow different applications to interoperate and work in a more personalized and intelligent way (GUARINO, 1998).

For achieving the objectives of an ontology-based system, it is however paramount that its supporting ontology is well-structured and developed in a coherent and consistent manner. Otherwise, the vocabulary may lead to misleading interpretations and the taxonomy may undermine information retrieval, thus disabling effective knowledge sharing and reuse. This can be rather disastrous. Once again, the need for a systematic methodology for ontology engineering becomes clear.

4.3 The Relation Between Ontology Engineering and Requirements Engineering

Grüninger and Fox (GRÜNINGER; FOX, 1995) claim that defining competency questions is a way to determine the scope of the ontology and outline a list of questions that the ontology will be able to respond. Similarly, requirements engineering aims at the definition of the behavior of the software system, considering functional and non-functional requirements.

Besides the already cited Boehm, other researchers have commented on the importance of these initial stages of software development. Lutz (LUTZ, 1993), for instance, found that the inadequate definition of requirements is responsible for a significant portion of errors detected during the process of systems development, especially in the case of dedicated systems, real-time and critics. And Davis (DAVIS, 1993) claims that eliminating engineering errors becomes increasingly difficult and expensive as the system progresses to later stages of their life cycle, such as the implementation stage.

To the extent of our knowledge, there is no study on the impact of a poor definition of the competency questions on the success of the ontology. However, we would expect it to be no different than what the requirement engineering researchers have found. Imagine, for instance that an ontology-based system has been developed for a travel agency. After the system is already running, the owner tries to run the following query: "What are the touristic attractions of Rio de Janeiro this week"? However, suppose that the only competency question regarding touristic attractions, defined before modeling the ontology was "What are the touristic attractions of a particular city?"Since such question did not take any temporal aspect into account, the travel agency owner's query cannot obtain the answer to his query.

This exemplifies how an incomplete competency questions definition can undermine the success of an ontology, as the basis of a software system. More complex examples could be conceived, including examples in which such information loss could seriously harm system interoperability or even blur the communication within organizations (without considering any automation whatsoever).

Although the importance of this definition step, within the field of ontology engineering, the existing methodologies make little use of elicitation and modeling techniques. And moreover, these methodologies still do not satisfactorily consider guidelines for the identification of concepts and relations (CAMPOS, 2010). Consequently, the ontology engineer remains so far with very limited resources to capture the requirements of the ontology in the process of ontology engineering.

4.4 Related Work: Competency Questions Definition Support in Existing Ontology Engineering Methodologies

Soares (SOARES, 2009) presents an analytical study which includes a list of 30 methodologies to build ontologies. For example, study selected in this work eight well-known methodologies for building ontologies.

To understand how each method deals with requirements engineering, some of the results were extracted and shown in Table 4.1 , which is then followed by a brief description of each methodology.

Table 4.1: Methodology vs. Technique for the definition of competency questions

Methodology	**CQ Definition Technique**
Sensus	Absent
Cyc	Manual extraction of the knowledge required
Grüninger and Fox	Informal Competency Questions
Uschold and King	Determine the purpose of the ontology
Kactus	List of requirements
Methontology	Defining the scope of the ontology
Method 101	Informal Competency Questions
Neon Methodology	Competency Questions in natural language

The Sensus methodology has no technique for determining competency questions. Following the methodology steps, in order to define the requirements, one should identify key terms of the domain and manually link these terms to the SENSUS ontology.

Cyc is a large common sense knowledge base, in which the first step to extract common sense knowledge is a manual extraction analyzing books, journals and articles, identifying issues that could be answered after reading these texts (LENAT et al., 1990)(LENAT, 1995). The reason why the Cyc knowledge base can be considered an ontology is because it can be useful as substrate for building different intelligent systems and also as a base for their communication and interoperability (GÓMES-PÉREZ; FERNÁNDES-LÓPEZ; CORCHO, 2004).

Michael Grüninger and Mark Fox in 1995 (GRÜNINGER; FOX, 1995) presented the methodology used for elaboration of the TOVE ontology. The first two steps were defined: developing motivation scenarios, which aim to identify problems in the current environment and specify informal competency question. These informal competency questions specify natural language requirements that the ontology should be able to attend. The example of an informal competency question could be: "Considering a young traveler, with limited budget for accommodation, what kinds of accommodations would be available?"(FOX, 1992).

Uschold and King 1995 consider as a first step to identify the purpose of the ontology under development, which aims at identifying: a) the need for development; b) the intended uses of the ontology; c) the relevant terms of the domain, For example, considering an ontology in the domain of interest of a travel agency, these terms could be: places, types of places, accommodation, types of accommodation, trains, buses, etc. (USCHOLD; KING, 1995)(USCHOLD et al., 1998).

Kactus (BERNARAS; LARESGOITI; CORERA, 1996), the first step would be to develop a list of needs or requirements that must be met by the application. Again, using a travel agency as example, an example of requirement could be: "Get the most suitable place for the customer, etc."(GÓMES-PÉREZ; FERNÁNDES-LÓPEZ; CORCHO, 2004).

The Methontology (GÓMES-PÉREZ; FERNÁNDES-LÓPEZ; CORCHO, 2004) includes a set of development stages (specification, conceptualization, formalization, integration, implementation and maintenance), and techniques to accomplish the planning, development and support of ontologies. In the specification activity, the following items should be defined: the purpose of ontology (what is intended), its usage scenario, scope and who the end-users are. This can be done by generating a document in natural language (GUIZZARDI; LUDERMIR; SONA, 2007).

The Method 101 proposes, as the first step, to determine the domain and scope of the ontology, answering questions such as: "What will we use the ontology for?", "What are the questions to which the ontology should provide answers?"(NOY; GUINNESS, 2001).

The Neon methodology proposes the following tasks for ontology specification: Identify purpose, scope and level of formality, Identify intended users, Identify intended uses, Identify requirements, Group requirements, Validate the set of requirements, Prioritize requirements and Extract terminology and its frequency (GÓMEZ-PÉREZ; SUÁREZ-FIGUEROA, 2009)(SUÁREZ-FIGUEROA et al., 2008). The goal of Identify requirements task is to obtain the set of requirements or needs that the ontology should fulfill. Users, domain experts and the ontology development team carry out this task taking as input a set of ontological needs for identifying the ontology requirements, using techniques as writing the requirements in natural language in the form of the competency questions. The output of this task is a list of competency questions written in Natural Language and a set of answers for the competency questions (SUÁREZ-FIGUEROA et al., 2008).

4.5 Goal-Based Modeling Method to Capture and Model Competency Questions

This section presenting our contribution to support competency questions via goal modeling. In subsection 4.5.1, we justify the use of Tropos in our work; subsection 4.5.2 summarizes our proposed Ontology Engineering approach, setting the context for the work presented in this dissertation; and subsection 4.5.3 explains in detail how we apply Tropos to capture and model competency questions.

4.5.1 Applying Tropos

Analyzing the most common methods found in the literature for the development of ontologies and their approaches to specify competency questions (see Table 4.1), we can identify three main objectives:

i defining the scope of the ontology, i.e. what parts of a specific domain the ontology should cover;

ii deciding the ontology's applicability or in other words, what the ontology will be actually used for;

iii specifying what questions the ontology must answer, i.e. what inferences the ontology should allow.

At this point, we believe that we can successfully apply Tropos (BRESCIANI et al., 2004) to achieve these goals. The following paragraphs justify our choice.

Tropos is an agent-oriented software engineering methodology which is founded on the concepts of actor and goal. Tropos's early requirements activity is concerned with understanding the organizational setting as it is, before a solution is actually conceived. The output of this activity is an organizational model, which includes relevant actors, their goals and interdependencies. By focusing on the problems which an organization aims at solving, the early requirements model provides us with the right context for the definition of the scope of the ontology (objective i). Moreover, by modeling actors, along with their goals and interdependencies, we gradually refine the model, up to a point in which the ontology's applicability becomes clear (objective ii)

In Tropos's late requirements activity, a solution is posed, generally in the form of a multiagent system. Still in this stage, the system is modeled in its operational environment by describing its most important functions. For that, the system under development is also represented by an actor who has a number of dependencies with other actors in the organization. These dependencies define the functional and nonfunctional requirements of the system. Such requirements are directly linked to the aforementioned objective iii, as it defines in detail what kind of support a system (in our case, an ontology-based system) should provide.

Finally, modeling resources provides us with a suitable concept for information modeling. Thus we can directly apply a Tropos concept to visually model competency questions and later, some initial concepts of the ontology.

To some up, the systematic analysis of goals provided by Tropos gradually leads us to understand what are the objectives involved in an organizational environment. And from this, we can understand that to satisfy these goals we need the right kind of information, so we can define the competency question based on this information need.

4.5.2 Proposed Approach

Figure 4.1 depicts the proposed Ontology Engineering approach. In early requirements, the goals of actor A1 of a given organization are analyzed, following the analysis methods advocated by Tropos. In late requirements, during this activity, competency questions are defined to accomplish the goals of this actor, when a solution is finally conceived in the form of an ontology-based system, dependencies are established between the system actor A2 and actor A1, the former acquiring competency questions CP1 and CP2 from the latter.

Besides modeling the competency questions, it is useful to elicit a few concepts, which directly derive from such questions. These concepts will later compose the actual ontology model. So still in late requirements, initial concepts of the ontology (here illustrated by c1, in diagram of actor A2) are captured and linked to the competency questions. In the ontology modeling activity, such initial concepts are specified and refined, applying a different conceptual modeling language, such as UML or ER diagram, for example.

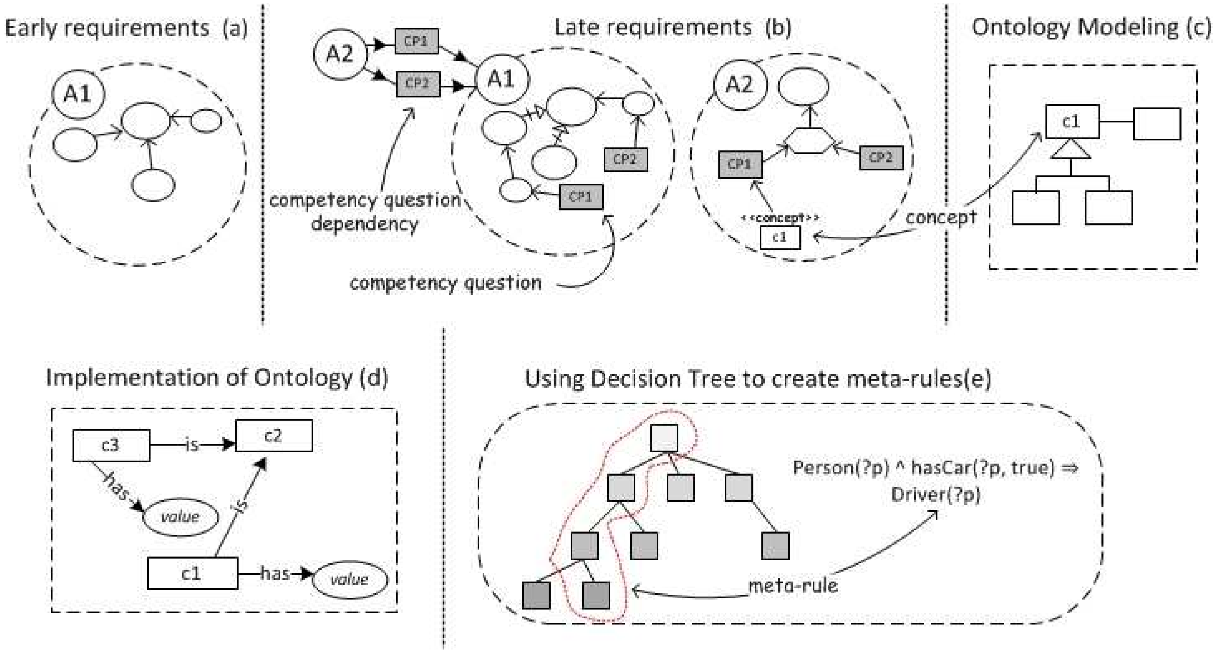


Figure 4.1: Proposed Ontology Engineering Approach.

The ontology created in the ontology modeling activity, can be used both as a conceptual reference model (design time), as a model to be used computationally (runtime) allowing, for example, the use of an inference engine. Furthermore, the ontology (design time) could also be translated into another model, for instance, the database model.

It is important to highlight that the scheme showed in Figure 4.1 goes beyond the ontology engineering process. It also comprehends steps (d) and (e), which deal with system development, and are domain specific steps, ie, were used in this study approach with the implementation of the ontology (d) and the use of decision tree for creating meta-rules (e). This does not mean that these steps should always be applied.

In our study presented here, the ontology created will be used computationally, in the implementation phase (d), we use the Resource Description Framework (RDF) to create a computation model. As the RDF model is only composed of triples (subject, predicate and object), we add some rules to the RDF model to support the conceptual model, detect of inconsistencies and allow inferences.

As a further step for our case study, we use decision tree to identify what we call meta-rules (e). Decision tree is used to learn patterns of epidemiological information from the territory of health. These meta-rules, are obtained after the decision tree is used as a tool for analyzing the suite of related data, available for our case study.

Moreover, the meta-rules do not tend to be static, such as the conventional rules. As the meta-rules are obtained through the results of the decision tree, they can change as the epidemiological profile changes, this is an important feature of the meta-rule not be static to better contribute with the quality of epidemiological study. For example, these meta-rules enable us to identify how a health condition is characterized in a given population, even if their characteristics change with time, the meta-rules will be updated, by the new results of the decision tree, it is also able to identify new health conditions in this population. This type of information is important for qualifying and enhancing the epidemiologic study.

It is important to state that in the ontology modeling activity, we apply the UFO foundational ontology (GUIZZARDI, 2005) to guide us on the ontology development process. Foundational ontologies have been recently recognized as important tools to guarantee semantic coherence and consistency when developing domain ontologies.

The difference between a foundational ontology and a domain ontology has to do with purpose and focus. While the latter is aimed at capturing a particular domain of discourse, the former is supposed to be domain-independent, dealing with general definitions such as entity, part-whole relation, and so on. In general, applying UFO provides us with safe guidance when deciding, for instance, if a concept should be seen as an entity or a property, also assisting us in finding out how two entities should be properly related. A deeper discussion on this topic, along with a detailed description of the UFO ontology is out of the scope of this chapter and can be found in (GUIZZARDI, 2005)(GUIZZARDI et al., 2009).

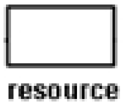
In this chapter, we discuss the phases (a), (b) and (c) while (d) and (e) are discussed in Chapter 5.

4.5.3 The applied Tropos constructs to competency questions

To model the information requirements, which in our case, are the competency questions, Tropos uses the constructs presented in Table 4.2. The second column of this table specifies how we apply such constructs.

The Tropos's resource construct (row A on Table 4.2) is naturally used to identify competency questions, since they are described as representing "a physical or an informational entity"((BRESCIANI et al., 2004), p.207). A competency question is clearly an informational entity and in our case, such entity is a resource to achieve the organizational actors goals and

Table 4.2: Tropos Constructs to capture the competency questions

	Tropos element	**Proposed use**
A		Used to identify the competency questions.
B		Extension of the resource construct to model the concepts involved in competency question. These concepts are the initial concepts composing the ontology.
C		Used to link competency questions to goals and concepts to competency questions.
D		Used to represent the dependence of the competence question.

later the ontology-based system goals. To model that, we connect these resources to goals via a means-end link (column C on Table 4.2). This means that the necessary information for the goal to be achieved is provided by the resource that represents the competency question. Depending on the complexity of the goal, including its information needs, several competency questions may be linked to the same goal.

As stated in subsection Proposed Approach, we propose some of the initial concepts composing the ontology are captured during the late requirements activity. To model such concepts in Tropos, we propose an extension, annotating these special kinds of resources as «concept», applying UML stereotypes (row B on 4.2). Please note that this element does not provide the regular semantics of a resource, since it would be unnatural to consider them as resources to achieve goals or plans. This justifies our extension. Something else we must care for is how to relate these concepts with the other Tropos's constructs. In this case, concepts can be only connected to competency questions, also via means-end link (row C on table 4.2). We find this a natural choice, since by the means of modeling specific concepts in the ontology; we are able to infer knowledge, thus answering a specific competency question.

A dependency relation (row D on Table 4.2) is applied to represent that an actor A2 depends on an actor A1 to acquire a given competency question which will later guide the ontology engineer in the ontology modeling activity. In this case, the dependence is a resource representing such competency question.

In terms of the process, there is an important observation to be made. In some cases, we know in advance that the system under development will be based on ontologies, even before starting the Tropos modeling activity. In this case, the competency questions are raised still in the perspective of the organization's actors. And later, such questions are delegated to the system actor. However, we also consider the case in which only in late requirements, a decision is made to deliver an ontology-based system. In this case, perhaps many of the system's goals, tasks and resources have already been elicited. However, it is important that the early requirements model is then revisited and competency questions modeling are normally performed.

4.5.4 Late Requirements: Modeling the Competency Questions

In this specific case study, we already started the work with the objective of developing an ontology-based system. So, the next step in our modeling process is to capture and model the Competency Questions to determine the scope of the ontology. The scope should define all the knowledge that should be in the ontology, as well as those that should not. The latter means that a concept should not be included, if there is not a competency question that uses it. This rule is also used to determine whether an axiom in the ontology should be included or not (BRUSA; CALIUSCO; CHIOTTI, 2006).

Figure 4.2 updates the model of Figure 3.4 (page 47) to include some competency questions that will later guide the tasks of the ontology engineer. As the reader may note, such competency question's start being captured in the perspective of the organizational actor, in this case the **Community Health Agent**.

Table 4.3 textually presents the competency questions. Please note that in the model, for reasons of space, we just use codes, such as CQ1, CQ2 etc. to represent such questions. That is why to understand the model, the analyst generally needs this auxiliary table.

Table 4.3: Competency Questions

CQ1. What are the patient with a given health condition that belong to families of high risk?
CQ2. What are the living conditions and health conditions of members from a family?
CQ3. What are the women by stage of life (teen, young, so on) pregnant?
CQ4. What are the women by stage of life (teen, young, so on) pregnant with some health conditions (hypertension, diabetes, so on)?
CQ5. Who are the patient that have no hypertension, but have the epidemiologic profile (eg, age, sex) of a patient with hypertension?
CQ6. What are the elderly who live alone?
CQ7. What are the families who did not receive visits in the month?

For *"know patients by health condition that belong to high risk families"*, with the purpose of perform monitoring to improve the health of these people the **Community Health Agent** must have sufficient information of what are the families at risk, and what are the health conditions of its members.

For example, knowing that a family consists of only one elderly member with hypertension and diabetes is important because it defines this family as having a need for special attention. Because an error in treatment, such as forgetting to take the medication, can cause more serious problems.

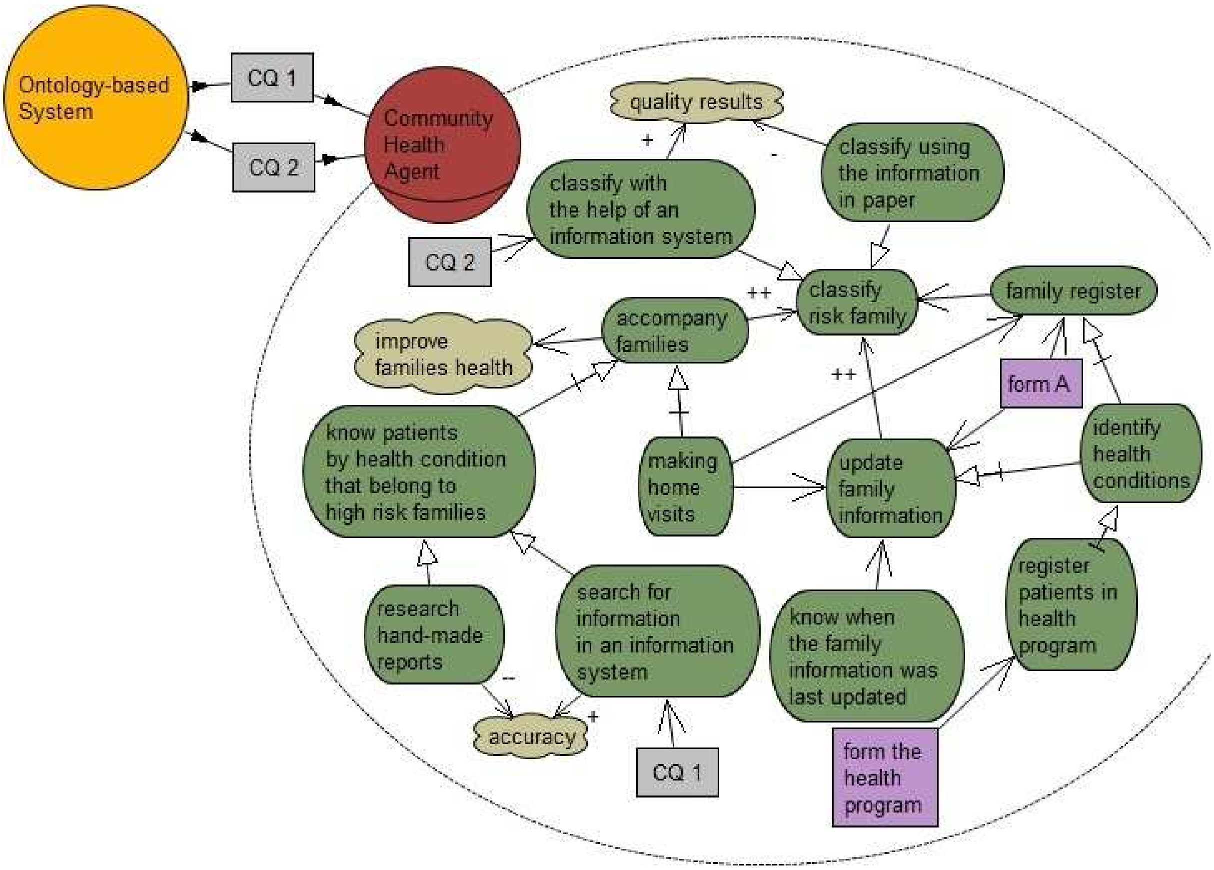


Figure 4.2: Late Requirements model capturing and delegating the competency questions the Community Health Agent.

To identify which are the high-risk families and the health condition of its members the **Community Health Agent** could *"research hand-made reports"* or *"search for information in an information system"* that will provide this support.

When choosing to research on hand-made reports, the **Community Health Agent** is increasing inaccuracy of the results, because reports on paper only turning information management more difficult increasing the possibility of errors. Errors that could, for example, identify incorrectly high-risk families, such as low-risk, leading to this family identified incorrectly serious problems. The fact of not properly performing the monitoring of a family, implies that only when something serious happens to a family member that the error is identified, in other words, would expect the problem to happen for later to be resolved, which is exactly contrary to the Family Health Strategy, where the main idea is to promote the health of the population, not waiting for the problem to occur in order to take action.

On the other hand, when the **Community Health Agent** uses an information system to get information on high-risk families and the health condition of its members, he is increasing the accuracy of the results because the information system can quickly provide the information in different ways, facilitating the study by the **Community Health Agent**.

At this point ***CQ 1*** assists the **Community Health Agent** in its activity by providing necessary information for the purpose of *"accompany families"* and also for this reason that the ***CQ 1*** is linked to the goal *"search for information in an information system"*.

To support the achievement of *"classify risk family"*, competency questions ***CQ 2*** assist the **Community Health Agent**, by providing information on housing conditions and health conditions of members of each family under their responsibility.

Still in the late requirements activity, we must model the system-to-be and its main goals. Figure 4.2 shows the Ontology-based System actor, representing this system. And it also shows some resource dependencies between the system and the **Community Health Agent**, in this case the dependencies are ***CQ 1*** and ***CQ 2*** only. These resources are the competency questions that should be answered by the system (objective iii of subsection 4.5.1).

Figure 4.3 updates the model of Figure 3.5 (page 48) to include some competency questions captured in the perspective of the organizational actor, in this case the **Nurse**.

The **Nurse** does not have as main activity, conducting home visits. One of its main activities is to supervise the activities, in this case the home visits, conducted by health community agents, their health team.

This is a very complex task, in our case study each health team has five (5) **Community Health Agent**. If, we consider an average of 11 home visits per agent per day, within twenty (20) working days in month we have a total of 1,100 (one thousand one hundred) monthly visits by the health team. How many patients will have their health information updated? How many patients or families may have changed their risk?

It is very difficult for **Nurses** to get a good monitor, without the aid of an information system to provide the necessary support to perform the task of supervision.

A major goal of the supervision of the **Nurse** is *"program families that should be visited"*, so it can be made that programming should be *"know the families by risk"*.

The first step in programming is to prioritize visits in accordance with the risk family. The next step is to know which patients are at risk these families as *"elderly who live alone"*, *"pregnant teenagers"* and *"young pregnant with hypertension"*. Here that ***CQ3***, ***CQ4*** and ***CQ6***, provide the information necessary for the **Nurse** to know quickly who are the patients at risk.

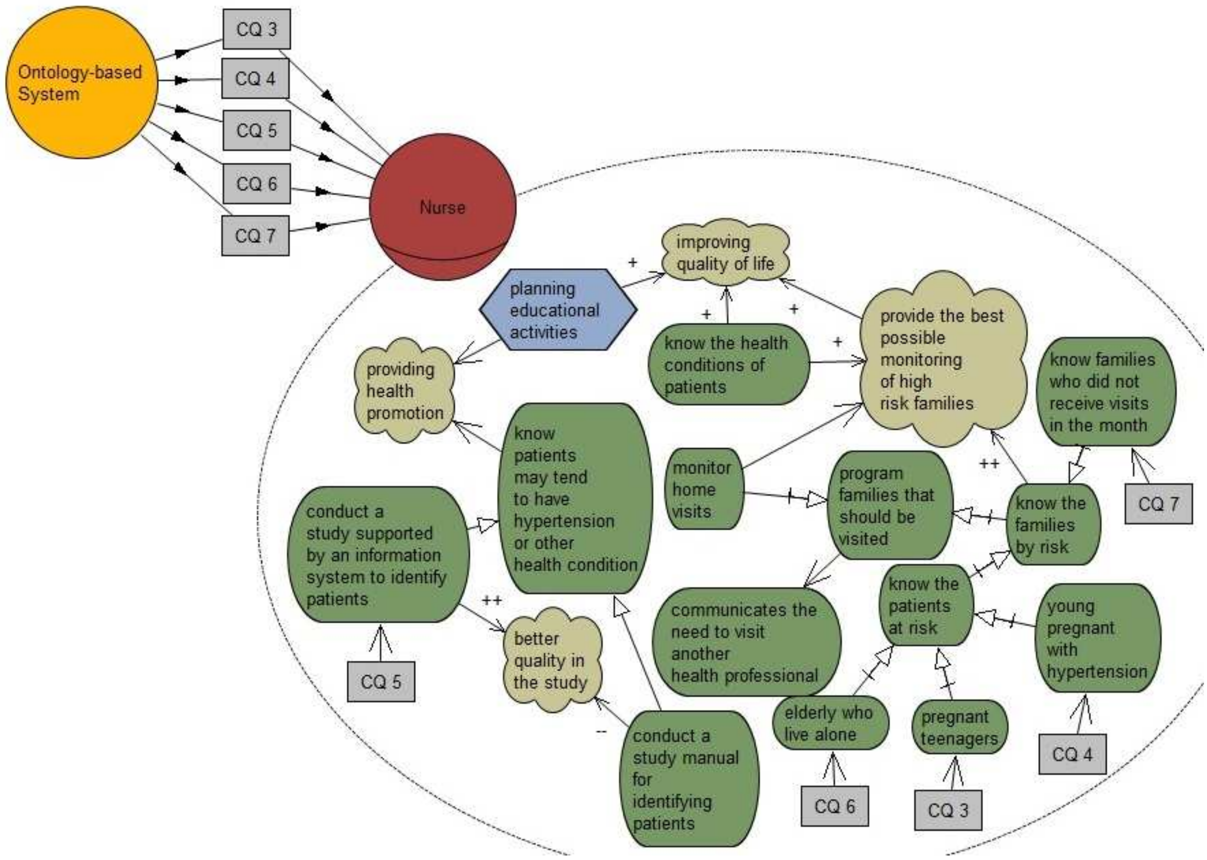


Figure 4.3: Late Requirements model capturing and delegating the competency questions the Nurse.

The advantage of the **Nurse** to have their such information provision, provides that he may be evaluating / knowing the patients at risk, weekly or even daily. With this monitoring nurses can prevent the patient stay without proper monitoring.

Another important objective is *"know families who did not receive visits in the month"*. Every family should be visited at least once a month. This goal is difficult to achieve for two main reasons: the number of families that have **Community Health Agent**, may make it impossible that all families are followed every month and also the fact that often do not find the resident (patient) at home.

Knowing this reality, ***CQ 7*** has as objective assist the **Nurse** for all families are monthly visits, especially families with medium and high risk. But even that is not possible to visit a family in the month, by ***CQ 7***, we can prevent it from happening again for several months.

Another difficult goal of the **Nurse**, also presented in section 3.3, is *"know patients may tend to have hypertension or other health condition"*.

The **Nurse** may, *"conduct a study supported by an information system to identify patients"* or *"conduct a study manual for identifying patients"*. When performing the study with the aid of an information system, the **Nurse**, contributes to *"better quality in the study"*.

The ***CQ 5***, provides the necessary information to the nurse so that this study can be performed. More details of how this study can be performed by an information system will be shown in the next chapter.

Still in the late requirements activity, we must model the system-to-be and its main goals. Figure 4.4 shows the **Ontology-based System** actor, representing this system. And it also shows the resource dependencies between the system and the **Community Health Agent** and **Nurse**. These resources are the competency questions that should be answered by the system (objective iii of subsection 4.5.1).

In Figure 4.4, we model the perspective of the Ontology-based System actor. It is important to note that although we are modeling the goals of the system, the competency questions refer to the ontology itself and not the system as a whole. Thus, the ontology is also represented as a resource within the system actor and a *means-end link* connects the competency questions to the ontology.

To accomplish the *"checking high-risk families"* goal, the system executes the *"get patient"*, this step could be, for example, retrieve patient information in a relational database or a file RDF (Resource Description Framework). The goal is then decomposed in two sub-goals: *"checking members of the family"* and *"checking health condition the patient"*.

At this stage we also identified some important concepts that should compose the ontology. Important concepts identified in ***CQ 1*** are: **Health Condition**, **Patient**, **Family** and **Risk**. These concepts are a special type of resource identified as **«concept»** which are connected to competency questions through a *means-end link*.

The idea here is not to be exhaustive, worrying about eliciting all concepts of the ontology. Capturing a few concepts that more obviously derive from the modeled competency questions provides a smooth transition to the following stage of ontology modeling.

It is also important to emphasize that it is not mandatory that every competency questions must generate a new concept to the ontology, as an example we have the ***CQ 4***. The ***CQ 4*** have important concepts that the ontology should have, but all have been identified in previous competency questions. An example would be: **Health Condition**.

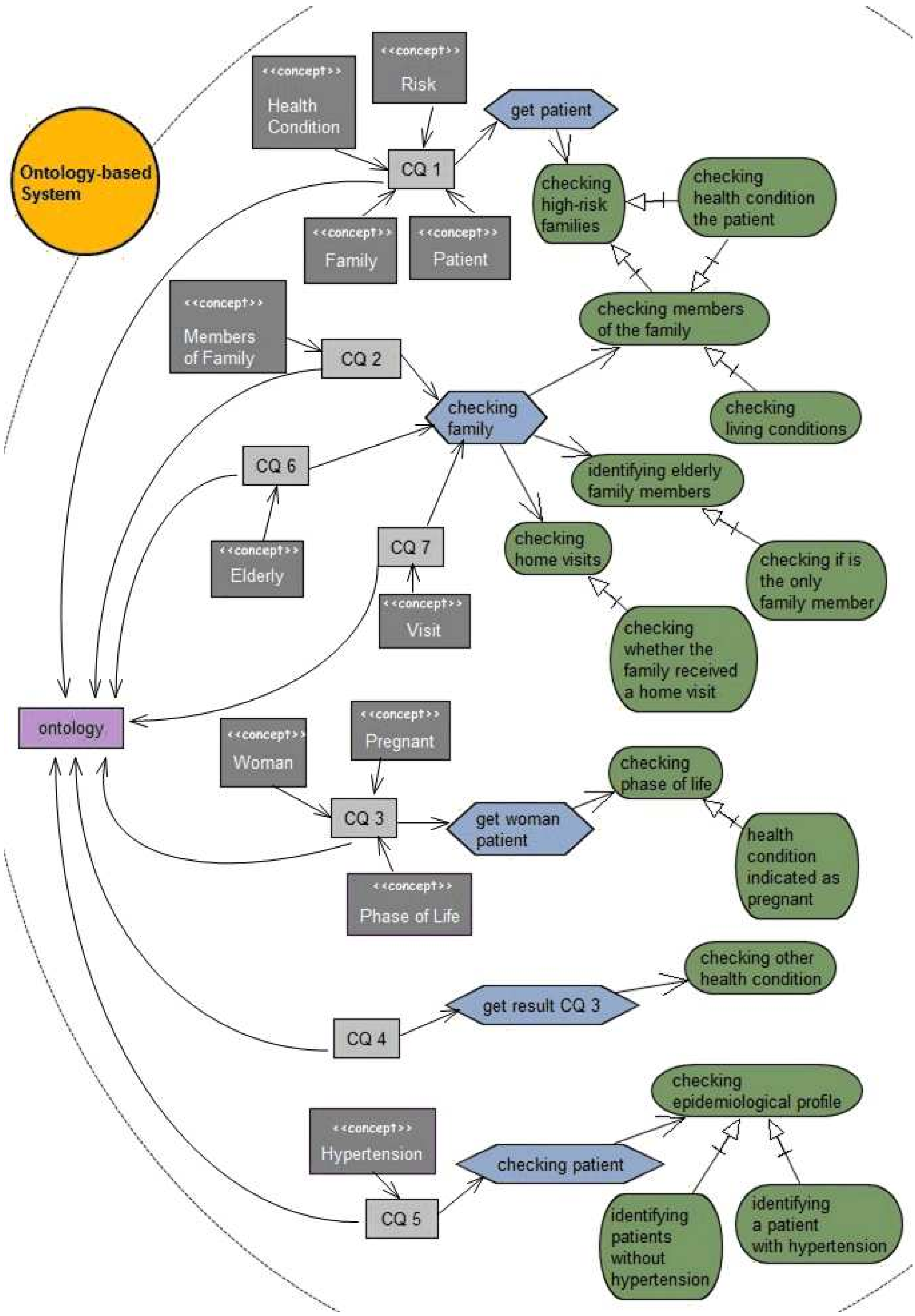


Figure 4.4: Actor diagram of Ontology-based system.

In ***CQ 4*** have also the fact that the competency questions is a complement of ***CQ 3***, in other words, ***CQ 4*** is equal to ***CQ 3*** increased restriction of the patient also having a health condition such as hypertension, diabetes, so on. For this reason that there is a plan *"get result CQ 3"*, ie, that ***CQ 4*** be answered by the ontology, simply, ***CQ 3*** run with the goal *"checking other health condition"* to be able to identify patients.

After capturing the competency questions, we are able to start eliciting a few concepts which will later compose the ontology that is the basis of the supporting system.

4.5.5 Concepts for ontology

To develop the ontology, we must refine the concepts coming from the late requirements model. In this dissertation, we modeled eleven concepts (see Figure 4.4): **Health Condition**, **Risk**, **Family**, **Patient**, **Members of Family**, **Elderly**, **Visit**, **Woman**, **Pregnant**, **Phase of Life** and **Hypertension**.

It is important to mention that not all concepts identified during the modeling phase, it will map directly to concepts of the ontology. For example, **Elderly**, is one **Phase of Life** as well as **Child**. **Hypertension** is a type of **Health Condition** and **Members of Family** are patients who belong to a **Family**. Although not directly mapped to the ontology, the ontology must be able to provide answers to these concepts, such as: What are the Members of a Family? or What are the Health Condition of the patient John?

Figure 4.5 shows a conceptual model of the ontology, showing how these two concepts may be refined. In this figure, we apply OntoUML (GUIZZARDI, 2005), an Ontologically Well-Founded version of UML for conceptual modeling and domain ontology engineering. This language includes a number of distinctions proposed by the Foundational Ontology UFO, some of which appear as stereotypes in the model of Figure 4.5.

Looking at Figure 4.5, we observe a number of finer-grained distinctions among the represented UML classifiers. For instance, by stereotyping **Visit** and **Person** as a *«Kind»*, the model indicates that these types provide a principle of identity for its instances and that the instances of those types are statically classified as instances of that type (e.g., an instance of **Person** cannot cease to be so without ceasing to exist). The stereotype of **Man** and **Woman** as *«Subkind»* has the same effect. In contrast, by representing **Patient**, **Community Health Agent** and **Nurse** as *«Role»* we have that every instance of those types can cease to be an instance of those types without ceasing to exist.

We also have a phase of life a **Person** can be partitioned into four stages according to their age, the model we have: **Child**, **Teenager**, **Young** and **Elderly** phase being represented by *«Phase»*. In this case, we have that every instance of those types can cease to be an instance of those types without ceasing to exist, for example, instance of a **Child** ceases to be an instance of

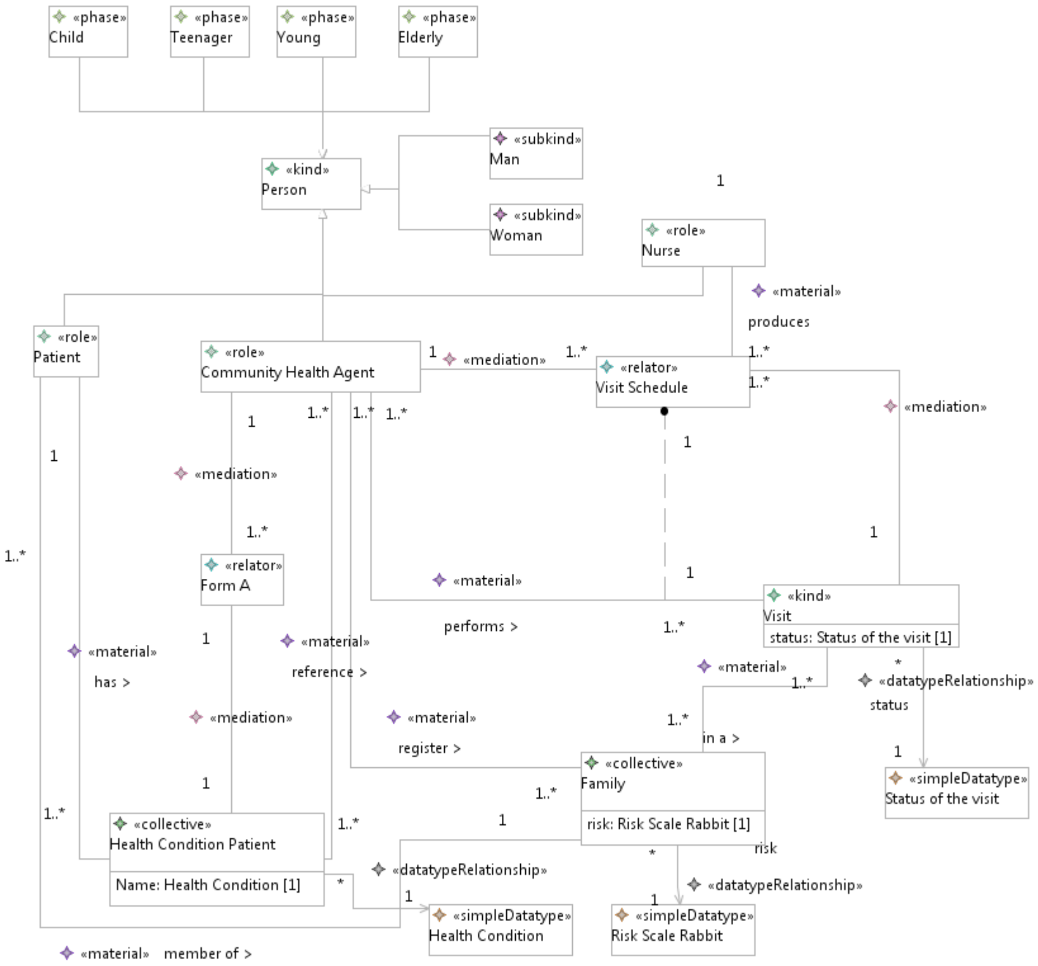


Figure 4.5: The Ontology using OntoUML.

Child when age is greater than 10 years, without ceasing to exist, in this case to be an instance of **Teenager**.

Moreover, a **Person** plays the role of **Patient** when being considered as part of a **Family** whose health risk is being assessed; and plays the role of **Community Health Agent** when mapping family health risks. In both cases, the conditions for instantiating those types are relational conditions. As a consequence, the same person can be both a patient and a health community agent. Furthermore, a family has a risk associated with it and that risk (a property) can be measured in a given Risk Conceptual Space (called Risk Scale Rabbit represented in OntoUML by a *«SimpleDatatype»*). In this model, this conceptual space is composed of four ordered datatype values: without risk, high, medium and low.

In the same way, the **Patient** may have a set of health conditions, represented in the model as *«Collective»* (**Health Condition Patient**). The **Health Condition** the **Patient** has is represented as *«SimpleDatatype»*, composed of values such as: Hypertension, Diabetes, Alcoholism, Pregnant and others.

In the model (Figure 4.5), we have the **Community Health Agent** performs **Visit**, but for this to occur it is necessary to the existence of one, **Visit Schedule**, which is produced by the **Nurse**. To ensure this relationship between the **Community Health Agent** and **Visit** (in this case a Home Visit), we have to, **Visit Schedule**, represented as *«Relator»* connecting the two in a *«material»* relation. Similarly, one might say, for example, that Mary works for the WHO [1] because there is an employment contract connecting them, and that John studies at the Federal University of Espirito Santo because there is an Enrollment relator connecting the two.

4.6 Conclusions

In recent years, we have been watching an explosion of works involving the use of Ontologies, mainly motivated by the need for semantic interoperability among domain models and applications; this is due to growing interest on the Semantic Web (GUIZZARDI et al., 2009). Benefits of an ontology-based system are, for example: a) reuse knowledge instead of reusing software; b) reuse and share application domain knowledge using a common vocabulary across heterogeneous software platforms and programming languages; and c) focus on domain structure, preventing them from being distracted by implementation details.

The ontology development begins with the definition of a set of questions, named competency questions. Competency questions is a way to determine the scope of the ontology and outline a list of questions that the ontology will be able to respond. Similarly, requirements engineering aims at the definition of the behavior of the software system, considering functional and non-functional requirements (GRÜNINGER; FOX, 1995).

At this point, we can draw a parallel between an ontology's competency questions and the requirements of a system. If we think carefully, both a competency question and a requirement identify a future characteristic of the ontology (the former) and the system to be (the latter). In software, this is then translated into functionality while in an ontology, this is materialized by the right set of concepts and relations, we can than conclude that the elicitation and modeling of competency questions should also undertake a thorough and precise process. Curiously, this is not the case so far. Most ontology engineering methodologies presuppose the existence of this set of questions and concentrate their efforts in defining the next stages of ontology development.

[1]WHO - World Health Organization

Goal modeling has been successfully applied in requirements engineering in the past years (GUIZZARDI; PERINI, 2005)(LIASKOS et al., 2010) to deliver software for organizations in general. One of the main strengths of this approach comes from the importance it gives to the understanding of the current state of the organization before a solution is actually conceived.

Our proposal, in an attempt to provide a solution to fill in the gap between the definition of the competency questions and the ontology modeling, encourages the use of goals, plans and resources to capture the competency questions that will contribute to achieving the organizational goals. From such competency questions, some initial concepts are drawn and, from then on, the ontology can go through iterative conceptual modeling stages.

An ontology provides a vocabulary this vocabulary comprehends logical statements that describe what the terms are and how they can or cannot be related to each other. Additionally, an ontology aims at specifying terms with unambiguous meanings, i.e. having their semantics independent of a particular reader and context. Otherwise, the vocabulary may lead to misleading interpretations and the taxonomy may undermine information retrieval, thus disabling effective knowledge sharing and reuse.

This can be rather disastrous, incomplete competency questions definition can undermine the success of an ontology, as the basis of a software system. More complex examples could be conceived, including examples in which such information loss could seriously harm system interoperability or even blur the communication within organizations.

Analyzing the most common methods found in the literature for the development of ontologies and their approaches to specify competency questions, we can identify that these methodologies still do not satisfactorily consider guidelines for the identification of concepts and relations. Consequently, the ontology engineer remains so far with very limited resources to capture the requirements of the ontology in the process of ontology engineering.

At this point, we believe that we can successfully apply Tropos (BRESCIANI et al., 2004) to achieve these goals. Tropos's early requirements activity is concerned with understanding the organizational setting as it is, before a solution is actually conceived.

The early requirements model provides us with the right context for the definition of the scope of the ontology. Moreover, by modeling actors, along with their goals and interdependencies, we gradually refine the model, up to a point in which the ontology's applicability becomes clear. In early requirements, the goals of actors of a given organization are analyzed, following the analysis methods advocated by Tropos.

In Tropos's late requirements activity, the system is modeled in its operational environment by describing its most important functions is also actor who has a number of dependencies with other actors in the organization. These dependencies define the functional and nonfunctional requirements of the system.

During this activity, competency questions are defined to accomplish the goals of this actor, when a solution is finally conceived in the form of an ontology-based system, dependencies are established between the system actors, the former acquiring competency questions, it is useful to elicit a few concepts, which directly derive from such questions.

These concepts will later compose the actual ontology model. So still in late requirements, initial concepts of the ontology are captured and linked to the competency questions. In the ontology modeling activity, such initial concepts are specified and refined, applying a different conceptual modeling language.

In our study, the ontology created will be used computationally, in the implementation phase, we use the Resource Description Framework (RDF) to create a computation model and also add some rules to the RDF model to support the conceptual model, detect of inconsistencies and allow inferences.

Furthermore, we use decision tree to identify what we call meta-rules. Decision tree is used to learn patterns of epidemiological information from the territory of health. These meta-rules, are obtained after the decision tree are used as a tool for analyzing the suite of related data, available for our case study.

In the next chapter (Chapter 5), we present an information system that can perform queries on the ontology created with the purpose of answer the competency questions, elicited in this chapter.

5 A system for query on the epidemiological profile

"Web users ultimately want to get at data quickly and easily ..."

Tim Berners-Lee

In this chapter we present the computational model of the ontology created using RDF, to allow query on the epidemiological profile.

5.1 Introduction

Given the results obtained in chapter 4, we present the system developed to allow performing queries on the epidemiological profile.

The computational model was created using RDF (Resource Description Framework), the choice of using RDF is given by its simplicity to describe a fact base. Besides RDF is the standard language used for providing data on the Internet through the linked-data (LINKED-DATA, 2012), allowing to create links of epidemiological information this work with other data sources available in the LOD (Linked Open Data) cloud.

This system uses the ontology created through the concepts identified in order to answer the competency questions (see Table 4.3). To give answers to these competency questions, the system uses the concepts of the ontology using the SPARQL language, query for RDF.

Besides the use of RDF to represent the ontology, rules were written using Euler language allowing the execution of inference and detection of inconsistencies in the ontology.

5.2 System Requirements

In this section, we will present the functional requirements that we could extract from the analysis presented in Chapters 3 and 4. We will then provide a conceptual model of an architecture that is capable of handling those requirements.

To identify the features that the system must possess it is essential first to find out what is really needed for a system with the purpose of allowing queries on the epidemiological profile of the territory. In this dissertation, the system should be able to answer the competency questions identified and modeled in subsection 4.5.4.

These competency questions provide the basic requirements for the system, but in addition to providing answers to competency questions, the system allows multiple queries to be submitted to the system by using the concepts defined in the ontology that is used in the system.

For example, ***CQ 1*** (What are the patient with a given health condition that belong to families of high risk?), wishes to know about the health conditions of patients in high risk families, the system must be able to provide this answer, but as shown in Figure 4.5 (OntoUML model), risk is an attribute of the **Family**, can be defined as: high, medium and low. Soon the system will also be able to answer questions involving families of medium and low risk. Another example is the use of inference, which we submit to the system in question: X has the ancestral Y ?.

The *ancestral* concept, is not defined in the created OntoUML model. In order to the system to be able to answer this question we have to define two rules, we can do this using Euler language, as shown in Figure 5.1.

Rule 1 - isRelative

```
{ ?x :isMemberOfFamily ?family. ?y :isMemberOfFamily ?family }
              => { ?x :isRelative ?y.  }.
```

Rule 2 - hasAncestral

```
{ ?x :isRelative ?y. ?x :hasAge ?ageX. ?y :hasAge ?ageY. ?ageY math:greaterThan ?ageX }
              => { ?x :hasAncestral ?y }.
```

Figure 5.1: Examples the rules using Euler.

The Rule 1 defines when two people are relatives. The rule defines that *x is relative of y*, if both are members of the same family. This first rule is necessary for us to define the next rule. The Rule 2 defines whether *x has ancestral y*. For this to be true, x should be relative to y, x and y must both have an age, and the age of y is greater than the age of x.

Adding these rules to the RDF model created, the system will reason about the concepts, relative and ancestral, being able to provide the answer to the question.

Another important requirement of the system is being able to perform an epidemiological study of a health condition, as a competency question ***CQ 5*** (Who are the patient that have no hypertension, but have the epidemiologic profile (eg, age, sex) of a patient with hypertension?). For the system to be able to answer this question, it is necessary for a tool to study and assess patients whose health condition hypertension and as age, sex and other characteristics of these patients influence the health condition. With this objective, this study used a decision tree (see section 2.7, page 38).

For these features to be provided by the system, the architecture was divided into three large blocks or layers. In Figure 5.2, it is shown how this architecture is organized, then explained each of these layers.

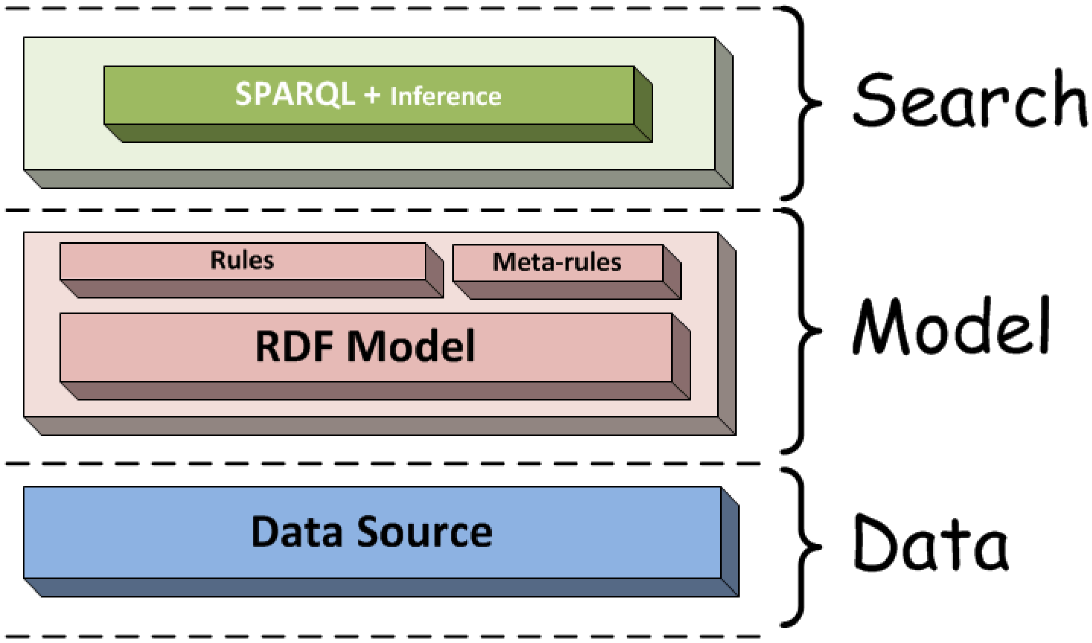


Figure 5.2: The 3-tier system architecture.

- **Data** - This layer is responsible for obtaining the data in its original form of storage, which may be, for example, a relational database. These data should be obtained and processed so that they can be passed to upper layer.

- **Model** - In this layer the computational model is dynamically created in RDF using the data passed by the lower layer. This was done with the help of software developed for this work that transforms the data received from the Data Source layer, into RDF triple. After you created the model, rules and meta-rules are also added to the model.

In this work we are considering such rules the concepts included in the model which were not obtained in an automated way using the decision tree, for example the concepts, *is-Parent* and *isAncestral* (see Figure 5.1, page 75). Meta-rules are concepts added to the model from the results obtained by the decision tree, for example the concept *hasEpidemiologicalProfile* (see Figure 5.8, page 83).

The rules are separated from the meta-rules, because the meta-rules can change over time, for example, a meta-rule that describes a person with the possibility of having diabetes can change the measure that epidemiological profile of the territory changes. When this occurs the meta-rule must be redone.

- **Search** - It is through this layer that the user interacts with the system. Writing SPARQL queries using existing concepts in the ontology, besides the concepts defined in the rules and meta-rules, the user can query information of his interest related to the epidemiological profile.

5.3 System Design

In this section we will present in detail how the model was developed in RDF and rules and meta-rules written in Euler were added to the model.

As computational support, we use the library *SemWeb.NET*(TAUBERER, 2012), which is a library that enables the interpretation of the technologies used for Semantic Web, such as RDF, through the use of SPARQL. This library was created for the C# programming language that is part of the Microsoft .NET Framework. The software created for this study was written using C# 4.0 and the current version 4.0 of the .NET Framework.

To perform the tests, we used real data from the basic health unit used in the case study (see section 3.2, page 42). These data were provided by the Department of Health, the city of Vitória-ES. The available data includes information relating to Form A, which contain information about the families registered by the Community Health Agent.

Information was provided in text file, then the data were read into a data structure in memory for later use in generating the RDF triple. An example of how the data was generated in RDF is shown in Figure 5.3.

In this example, the data is shown above a family and its members. Line 2 through line 12 shows the information concerning the family, for example, it is reported in line 4 that this family has no risk family (*rdf:resource="http://epi/Risk/WithoutRisk"*). In this example, the family has

```
1 <rdf:Description rdf:about="http://epi/MA_13_NF_263">
2     <ep:hasMicroArea>13</ep:hasMicroArea>
3     <ep:hasNumberOfFamily>263</ep:hasNumberOfFamily>
4     <ep:hasRisk rdf:resource="http://epi/Risk/WithoutRisk" />
5     <ep:hasWaterTreatment rdf:resource="http://epi/WaterTreatment/Filtration" />
6     <ep:hasWaterSupply rdf:resource="http://epi/WaterSupply/Public_Network" />
7     <ep:hasSewage rdf:resource="http://epi/Sewage/Public_Network" />
8     <ep:hasTargetWaste rdf:resource="http://epi/TargetWaste/Collected" />
9     <ep:hasInCaseOfDiseaseSearch rdf:resource="http://epi/InCaseOfDiseaseSearch/Hospital" />
10    <ep:hasModeOfTransportUsed rdf:resource="http://epi/ModeOfTransportUsed/Car" />
11    <ep:hasNumberOfRooms>9</ep:hasNumberOfRooms>
12    <ep:hasNeighborhood>PRAIA DO CANTO</ep:hasNeighborhood>
13    <ep:hasMembersOfFamily>
14        <rdf:Description rdf:about="http://epi/282805_GERALDO_[illegible]">
15            <ep:hasName>GERALDO [illegible]</ep:hasName>
16            <ep:hasAge>81</ep:hasAge>
17            <ep:hasSex>M</ep:hasSex>
18            <ep:hasRegistryNumber>282805</ep:hasRegistryNumber>
19        </rdf:Description>
20    </ep:hasMembersOfFamily>
21    <ep:hasMembersOfFamily>
22        <rdf:Description rdf:about="http://epi/282809_MARIA_[illegible]">
23            <ep:hasName>MARIA [illegible]</ep:hasName>
24            <ep:hasAge>86</ep:hasAge>
25            <ep:hasSex>F</ep:hasSex>
26            <ep:hasRegistryNumber>282809</ep:hasRegistryNumber>
27            <ep:hasHealthCondition rdf:resource="http://epi/HC/Hypertension" />
28        </rdf:Description>
29    </ep:hasMembersOfFamily>
30 </rdf:Description>
```

Figure 5.3: Example of a family modeled using RDF.

two members (Geraldo and Maria), the information regarding the family members are separated by tags *<ep:hasMembersOfFamily>*. In line 27, it is shown that this family member has a health condition, namely hypertension (*rdf:resource="http://epi/HC/Hypertension"*).

After this model is created, the RDF file is saved to disk. One only need to generate the file again if it is necessary to change the structure of the RDF model, such as the creation of a new triple. Because the model is dynamically created and then stored in an RDF file on disk, if new information, ie, data from families, patients, son on, should be added to the file, the model must be rebuilt.

In our study, the choice was to use this approach because it does not affect the study, in a production environment where it was necessary to use data up to date, we could use another tool such as Virtuoso (VIRTUOSO, 2012).

After the RDF model is created, it must be loaded with all the rules (including the meta-rules). Figure 5.4, is an example of how the model is loaded by the system. In line 7 the file *"model.rdf"* is loaded from disk into memory. Despite exceeding 60 MB in size, file reading is done quickly by *SemWeb.NET* library. In line 12, the rules are added to the model. These rules are written in Euler, as the example in Figure 5.1, page 75.

In section 5.2, we present the technique used to generate the meta-rules. Now we will show

```
MemoryStore store;

private void frmMain_Load(object sender, EventArgs e)
{
    store = new MemoryStore();

    using (RdfXmlReader reader = new RdfXmlReader("model.rdf"))
    {
        store.Import(reader);
    }

    store.AddReasoner(ReasonerController.GetAllRules());

}
```

Figure 5.4: Loading RDF model with the rules.

how decision trees were used to aid in the creation of meta-rules to the ontology, in order to assist in the epidemiological study.

For example, let us consider the ***CQ 5*** (Who are the patient that have no hypertension, but have the epidemiologic profile (eg, age, sex) of a patient with hypertension?). To answer this question, the first step is to understand the epidemiological profile of patients who have the health condition of hypertension.

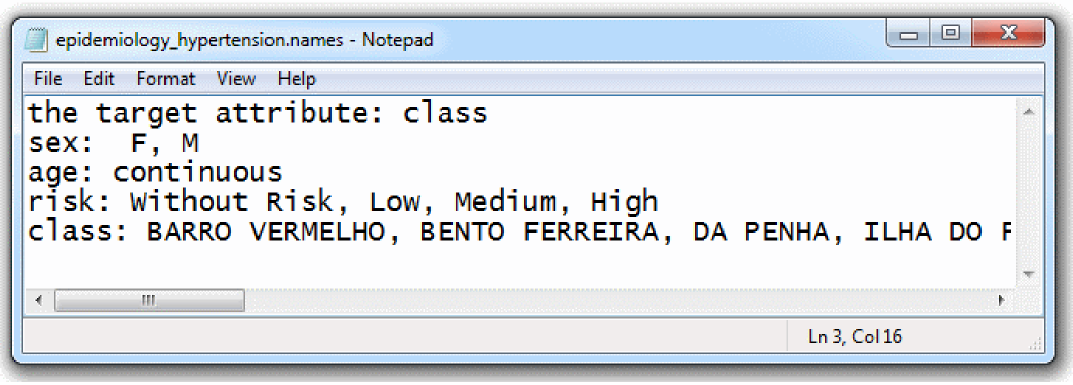
For that we are using as a tool decision tree. In this example, consider for all patients who have the health condition of hypertension the attributes: sex, age, family risk and neighborhood of residence. The goal is that after analysis using a decision tree, we can know for what values of such attributes, the patient tends to have hypertension.

For this analysis, we use the algorithm called C4.5 decision tree (QUINLAN, 1993) and we use the implementation in Java called FC4.5 (Fast C4.5).[1]

To execute the algorithm FC4.5, you need two configuration files. A file with the extension *.names*, which has the attributes and possible values for each of these attributes, example Figure 5.5. In this example the attribute that represents the neighborhood of residence was configured with the target attribute, so the values for this attribute must be configured using the *class* keyword.

The other file, Figure 5.6, with the extension *.data*, has what we call data for training. This file has sex, age, risk of family and neighborhood of each patient that has a health condition of hypertension. This is the set of data used by the algorithm to create the decision tree. In our

[1]Fast C4.5 - http://code.google.com/p/fc45/



epidemiology_hypertension.names - Notepad
File Edit Format View Help
```
the target attribute: class
sex:  F, M
age: continuous
risk: Without Risk, Low, Medium, High
class: BARRO VERMELHO, BENTO FERREIRA, DA PENHA, ILHA DO F
```
Ln 3, Col 16

Figure 5.5: Configuring file *.names* for decision tree.

example the file contains 2.391 lines.

epidemiology_hypertension.data - Notepad
File Edit Format View Help
```
F, 57, Without Risk, ITARARE
M, 66, Without Risk, ITARARE
M, 60, Without Risk, PRAIA DO CANTO
F, 55, Without Risk, PRAIA DO CANTO
M, 68, Medium, DA PENHA
M, 78, Without Risk, PRAIA DO CANTO
F, 75, Without Risk, PRAIA DO CANTO
F, 79, High, ITARARE
F, 52, High, ITARARE
F, 75, High, ITARARE
M, 78, High, ITARARE
F, 73, High, ITARARE
F, 63, High, ITARARE
M, 72, Without Risk, SANTA LUIZA
F, 68, Low, ITARARE
F, 55, Without Risk, SANTA LUIZA
F, 79, Without Risk, SANTA LUIZA
F, 76, Low, SANTA LUCIA
M, 62, Without Risk, BARRO VERMELHO
```
Ln 1, Col 1

Figure 5.6: Configuring file *.data* for decision tree.

Running the algorithm FC4.5 following configuration files, will result in the decision tree generated in this dissertation, the decision tree was exported to an html file in order to facilitate the visualization of results.

The results are always two decision trees, a tree without pruning and other tree with pruning. In our study, we used to generate the meta-rules the decision tree with pruning. Pruning is a technique in machine learning that reduces the size of decision trees by removing sections of the tree that provide little power to classify instances.

The decision tree with pruning result is shown in Figure 5.7. We can observe that the tree are does not have a great depth, which contributes to the creation of meta-rules that not so complex. In this example, we could create up to eight meta-rules, one for each leaf node.

The simplest meta-rule to be created would be: *every patient member of a family of high-risk in the neighborhood Itararé, tends to be hypertensive*.

It is very important to explain that the study of decision tree, can not be taken as an absolute truth. For this reason, we call meta-rules, not rules. The results obtained after the application of the decision tree algorithm, leads us to think about some features, referring to the data set that was used in training.

Using the meta-rule cited above, we may ask: Why is there this tendency of all high-risk families in the neighborhood Itararé, having members with hypertension?

It is through this question that the epidemiological study moves. We could trigger a series of analyzes to be performed as: What are the living conditions of these patients? Are there families in the neighborhood Itararé which do not have members with hypertension?

The study presented here was developed with this purpose, i.e., seeking to offer a set of tools that can support the study, analysis and understanding of the epidemiological profile of a given territory of health.

Figure 5.8 illustrates how meta-rules are written using Euler[2], what is the inference engine standard of *SemWeb.NET* library. In this example, two meta-rules are being defined.

Both meta-rules define the concept (in RDF represents the predicate) *hasEpidemiological-Profile*. The first meta-rule defines that patients who are family members of high-risk at the neighborhood Itararé, which does not have the health condition hypertension, have epidemiological profile of patients with hypertension.

The second meta-rule defines the same concept: *hasEpidemiologicalProfile*. What changes the conditions for the patient to be considered as epidemiological profile for hypertension. The meta-rule defines: patients aged less than 88 years (in the decision tree is displayed less than or equal to 87, this change was necessary because the library *SemWeb.NET* only contains command *math:lessThan*), who are members of a family of medium-risk the neighborhood Itararé, which does not have the health condition hypertension, have epidemiological profile of patients with hypertension.

[2]Euler Proof Mechanism - http://www.agfa.com/w3c/euler/

Figure 5.7: Results of decision tree pruning for patients with hypertension.

The other meta-rules could be created in the same way that these two are shown here. It is very important during the creation of meta-rules to prevent the generation of contradictions, this can lead to generation of incorrect inferences in his model, thus leading to invalid results during the consultations.

Elsewhere on the system where we use rules, was to define the phase of life the patient. In OntoUML model (see Figure 4.5, page 70), these entities are identified with the stereotype *Phase*.

The age ranges to define when a patient is: child, teenager, and so on, is shown in Table 5.1. These values are established by the Ministry of Health of Brazil.

Table 5.1: Phase of Life vs. Age Ranges

Phase of Life	Age Range
Child	*age* >= 0 and *age* < 11
Teenager	*age* > 10 and *age* < 20
Young	*age* > 19 and *age* < 60
Elderly	*age* > 59

To be able to perform a query informing phase of life of patients is necessary to define the concept *phaseOfLife*. This concept was defined by creating a rule for each phase of life (four rules in total), as shown in Figure 5.9.

Meta-Rules

```
{
    ?patient :isMemberOfFamily ?family.
    ?family :hasRisk <http://epi/Risk/High>.
    ?family :hasNeighborhood "ITARARE".
    ?patient :hasNoHealthCondition <http://epi/HC/Hypertension>.
}
=>
{
    ?patient :hasEpidemiologicalProfile <http://epi/HC/Hypertension>.
}.
----------------------------------------------------------------------
{
    ?patient :hasAge ?age.
    ?age math:lessThan 88.
    ?patient :isMemberOfFamily ?family.
    ?family :hasRisk <http://epi/Risk/Medium>.
    ?family :hasNeighborhood "ITARARE".
    ?patient :hasNoHealthCondition <http://epi/HC/Hypertension>.
}
=>
{
    ?patient :hasEpidemiologicalProfile <http://epi/HC/Hypertension>.
}.
```

Figure 5.8: Example meta-rules using Euler.

Child

```
{
    ?patient :hasAge ?x.
    ?x math:greaterThan -1. ?x math:lessThan 11
}
=>
{
    ?patient :phaseOfLife \"Child\"
}.
```

Young

```
{
    ?patient :hasAge ?x.
    ?x math:greaterThan 19. ?x math:lessThan 60
}
=>
{
    ?patient :phaseOfLife \"Young\"
}.
```

Teenager

```
{
    ?patient :hasAge ?x.
    ?x math:greaterThan 10. ?x math:lessThan 20
}
=>
{
    ?patient :phaseOfLife \"Teenager\"
}.
```

Elderly

```
{
    ?patient :hasAge ?x.
    ?x math:greaterThan 59
}
=>
{
    ?patient :phaseOfLife \"Elderly\"
}.
```

Figure 5.9: Example of rules for defining the phases of life.

Examples of how to perform queries using the meta-rules and rules are demonstrated in section 5.5.

5.4 Prototype

To verify the validity of the architecture proposed in 5.2 we built a small prototype that is supposed to execute the test cases described in the next section. The prototype consists of a desktop application written in C#.

The first step before the application is open is to load the RDF model to the meta-rules and rules created. After this point, we can perform queries using SPARQL.

The results of SPARQL queries, the library *SemWeb.NET* are returned in the XML standard in accordance with the recommendation of the W3C (SPARQL-XML, 2012). To make viewing of the most friendly, the result in XML was transformed into an HTML output using an XSL file.

In Figure 5.10 is shown performing a SPARQL query with the display its results. In this example, is being queried all patients who have the health condition of hypertension. With the purpose to facilitate queries was simplified how the component *object*, the RDF triple must be specified. In our example the reference to the *object* hypertension, was done only informing *HC/Hypertension*, and not the full path would be *http://epi/HC/Hypertension*.

This important feature unique URI in RDF was preserved, when you do not specify the full path of the *object* in the query, the system itself performs this work.

As a result of the query are displayed in 2.396 records that represent patients who have the health condition of hypertension.

5.5 Using the Prototype

In this section we will present how to use the prototype presented in the previous section to answer the competency questions presented in section 4.5.4 (see Table 4.3, page 63).

This section is organized in a subsection for each competency question to be answered by the system.

5.5.1 Answering the competency question - *CQ 1*

Through the ***CQ 1*** (What are the patient with a given health condition that belong to families of high risk?) can be identified by the health condition of patients at high risk family. Shown in Figure 5.11 is a SPARQL query can answer that question.

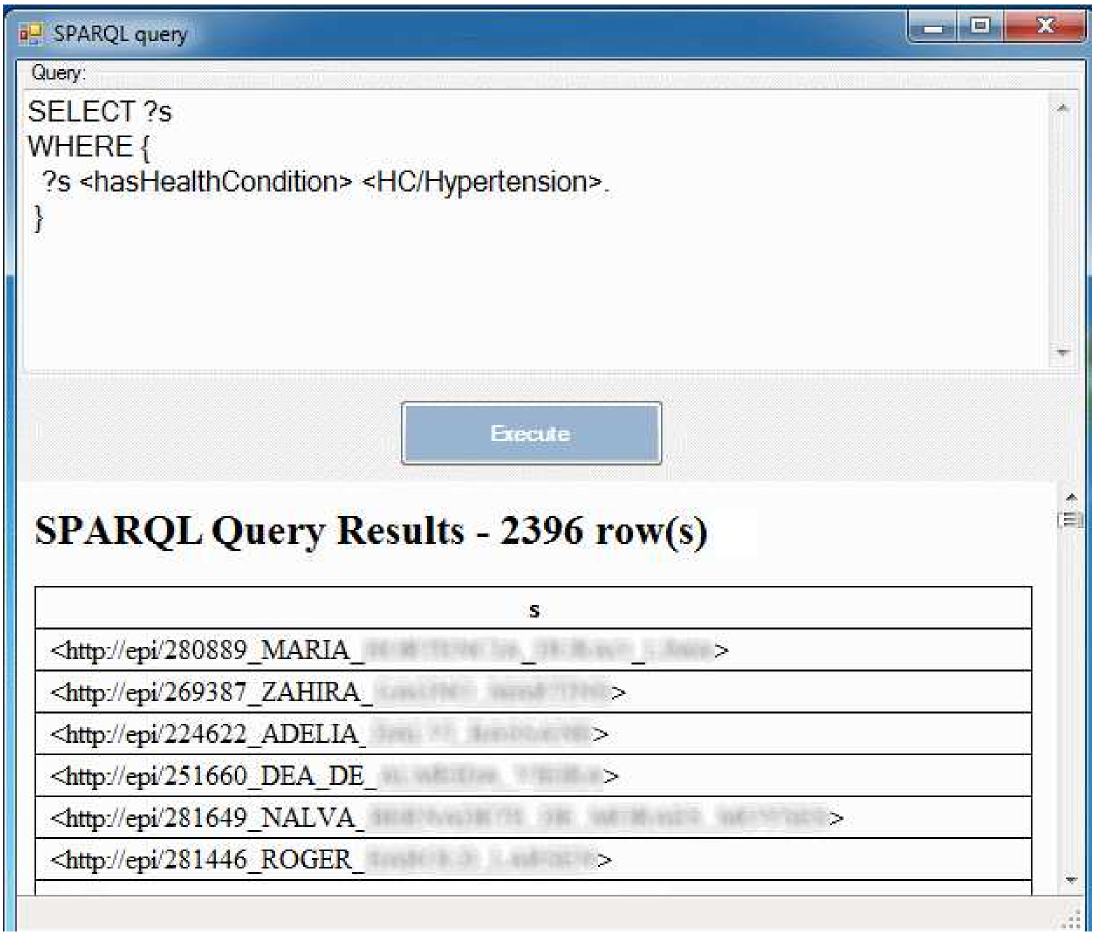


Figure 5.10: Example of how to query using the prototype created for testing.

In our example we are querying all patients having the health condition of diabetes (*?p <hasHealthCondition> <HC/Diabetes>*), are members of a family (*?f <hasMembersOfFamily> ?p*) of high risk (*?f <hasRisk> <Risk/High>*). In our example 17 patients are returned.

To query patients to the health condition hypertension, would suffice to replace *?p <hasHealthCondition> <HC/Diabetes>* by *?p <hasHealthCondition> <HC/hypertension>*. Thus ***CQ 1*** is fully met by the system.

5.5.2 Answering the competency question - *CQ 2*

Through the ***CQ 2*** (What are the living conditions and health conditions of members from a family?) we can know the living conditions and health conditions of family members.

For this example we split the answer into two SPARQL queries to facilitate understanding and for easy viewing of results. A query returns the living conditions of the family in another query are returned members of families with their health conditions.

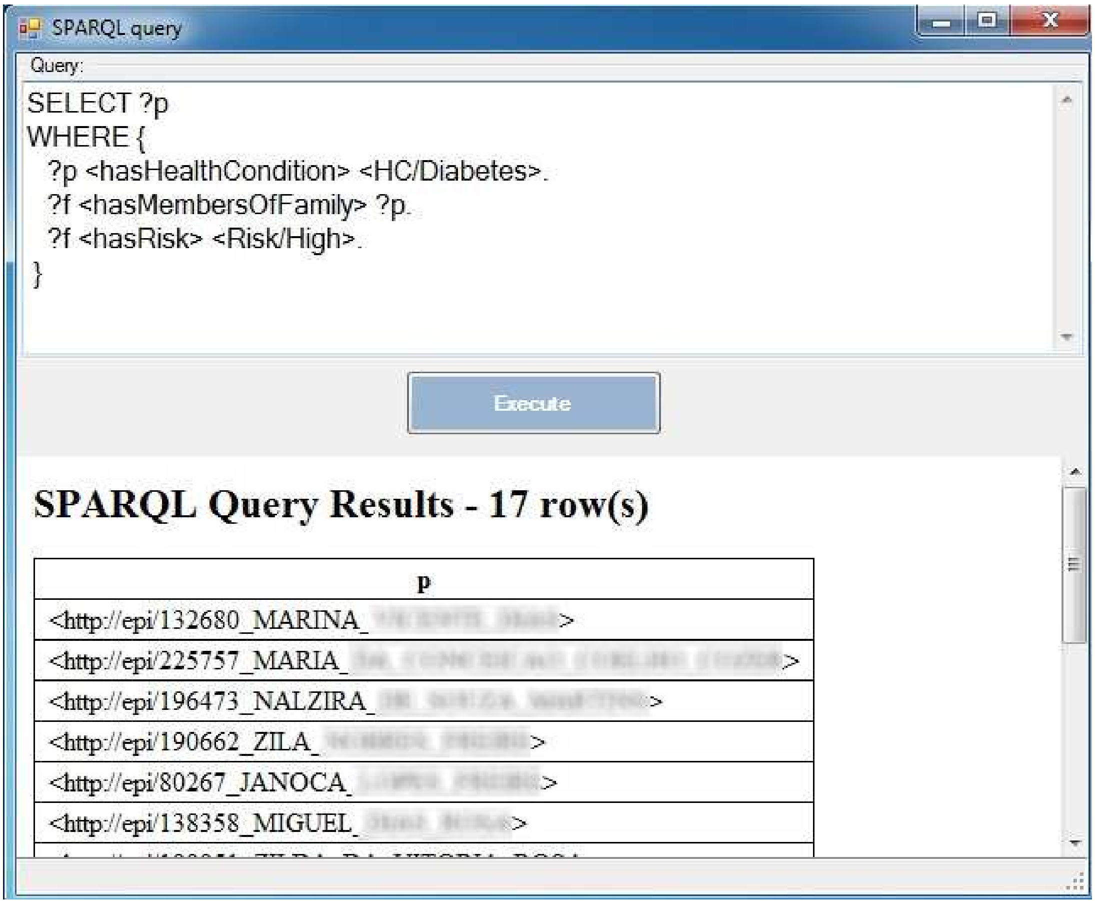


Figure 5.11: SPARQL to answer the competency question - ***CQ 1***.

Shown in Figure 5.12 is a SPARQL query that returns the living conditions of the family. For this example, we are referring to micro area 42 (<*hasMicroArea*>), family 42 (<*hasNumberOfFamily*>), to know the following living conditions of the family: type water treatment (<*hasWaterTreatment*>), if has power (<*hasElectricity*>), type water supply (<*hasWaterSupply*>) and the type of house (<*hasHouseHoldType*>).

In Figure 5.13 shows the SPARQL to return the health conditions of family members. For this example, family 42 (<*hasNumberOfFamily*>) belonging to micro area 42 (<*hasMicroArea*>), are returned all family members (<*hasMembersOfFamily*>), with the health condition of each of the members (<*hasHealthCondition*>).

As a result eight records are displayed. As three different people in this family. For each family member have: Aryane with 1 (one) health condition, Beatriz with two (2) health conditions and Edvaldo with five (5) health conditions, totaling 8 records returned by the query.

To see another family would be enough replace the micro area or family. In this example we use two SPARQL queries to obtain the desired response, but it is possible to unify the two queries without difficulty.

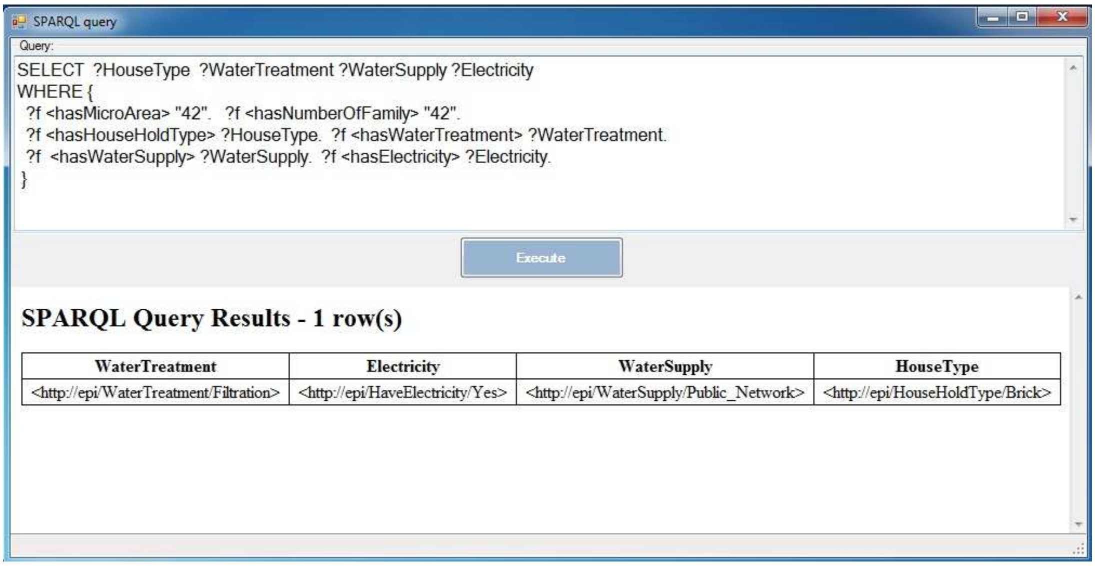


Figure 5.12: SPARQL to answer the competency question - ***CQ 2*** - living condition.

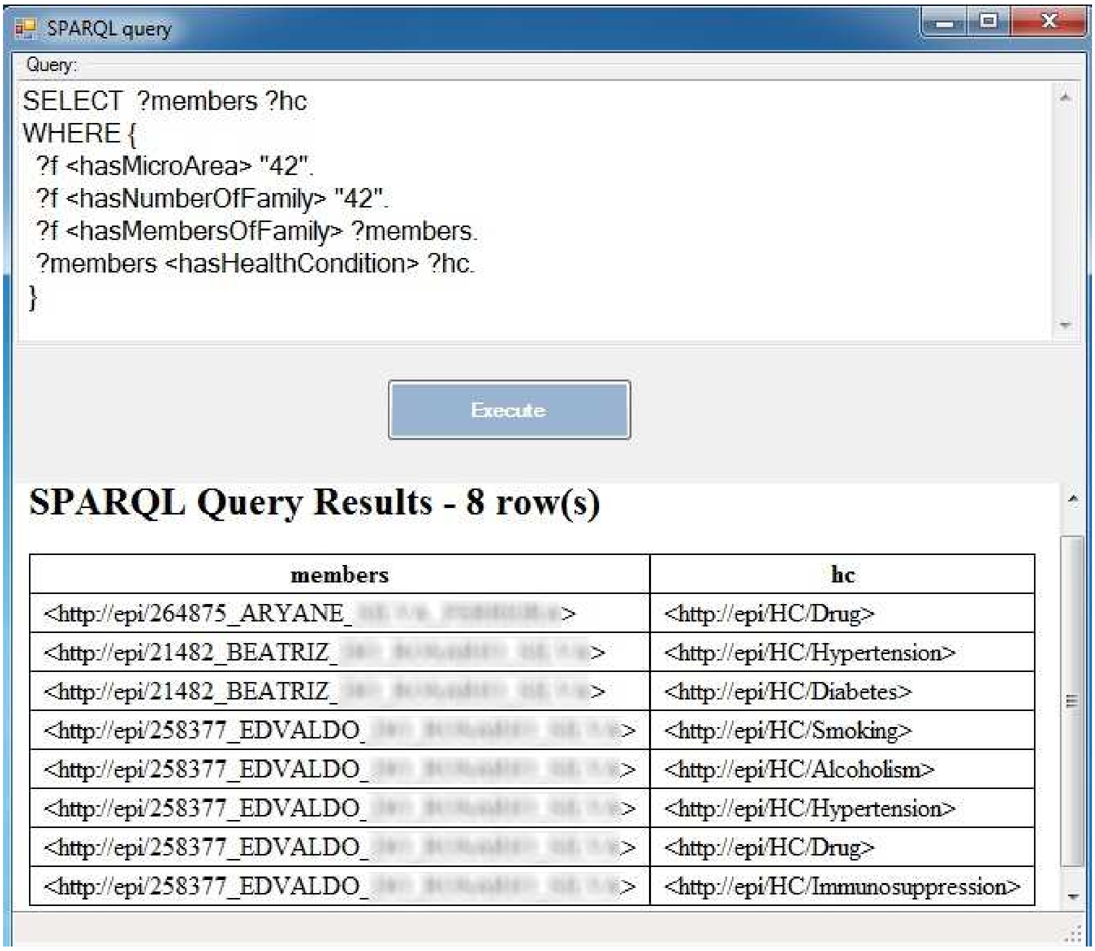


Figure 5.13: SPARQL to answer the competency question - ***CQ 2*** - health condition.

5.5.3 Answering the competency question - *CQ 3*

Through the ***CQ 3*** (What are the women by stage of life (teen, young, so on) pregnant?) we can know the vulnerable patients, ie patients at risk.

In Figure 5.14, we have one SPARQL query to answer ***CQ 3***. This query returns all teenager patients (*?p <phaseOfLife> "Teenager"*) who are in the health condition of gestation (*?p <hasHealthCondition> <HC/Gestation>*), or in other words, pregnant.

In our example we did not perform filtering by sex of the patient, because pregnancy is already restricted to female sex, but this could be done added the following RDF triple: *?p <hasSex> "F"*.

As a result, we have the name of nine women at the stage of teens who are pregnant. In addition to the names in this example also exhibit the age (*?p <hasAge> ?age*) of each of the patients in this situation.

We carry out this example to demonstrate the importance of one tool to aid health professionals in identifying patients at risk. We can identify the case where one patient, only 14 years old are pregnant.

Surely this pregnancy is a risk to the mother and the child, the sooner the patient is accompanied, more serious problems can be avoided.

5.5.4 Answering the competency question - *CQ 4*

On the competency question ***CQ 4***, we have one complement of ***CQ 3***. That is, the SPARQL query ***CQ 4*** is equal to the SPARQL query ***CQ 3***, just added the restriction of another health condition, in addition the health condition gestation.

In the example shown in Figure 5.15, the result of the following query: young pregnant women, who also has the health condition of diabetes (*?p <hasHealthCondition> <HC/Diabetes>*).

As a result we have only one patient (Carla) with 28 years of age who is pregnant and also has a health condition diabetes. To query other health condition would be sufficient replace *?p <hasHealthCondition> <HC/Diabetes>* by *?p <hasHealthCondition> <HC/Hypertension>*, and so on.

Very important addition to the triple RDF *?p <hasHealthCondition> <HC/Hypertension>* without replacement of the triple *?p <hasHealthCondition> <HC/Diabetes>*, mean that the patient must have both health conditions together, ie, the patient must be hypertension and diabetes.

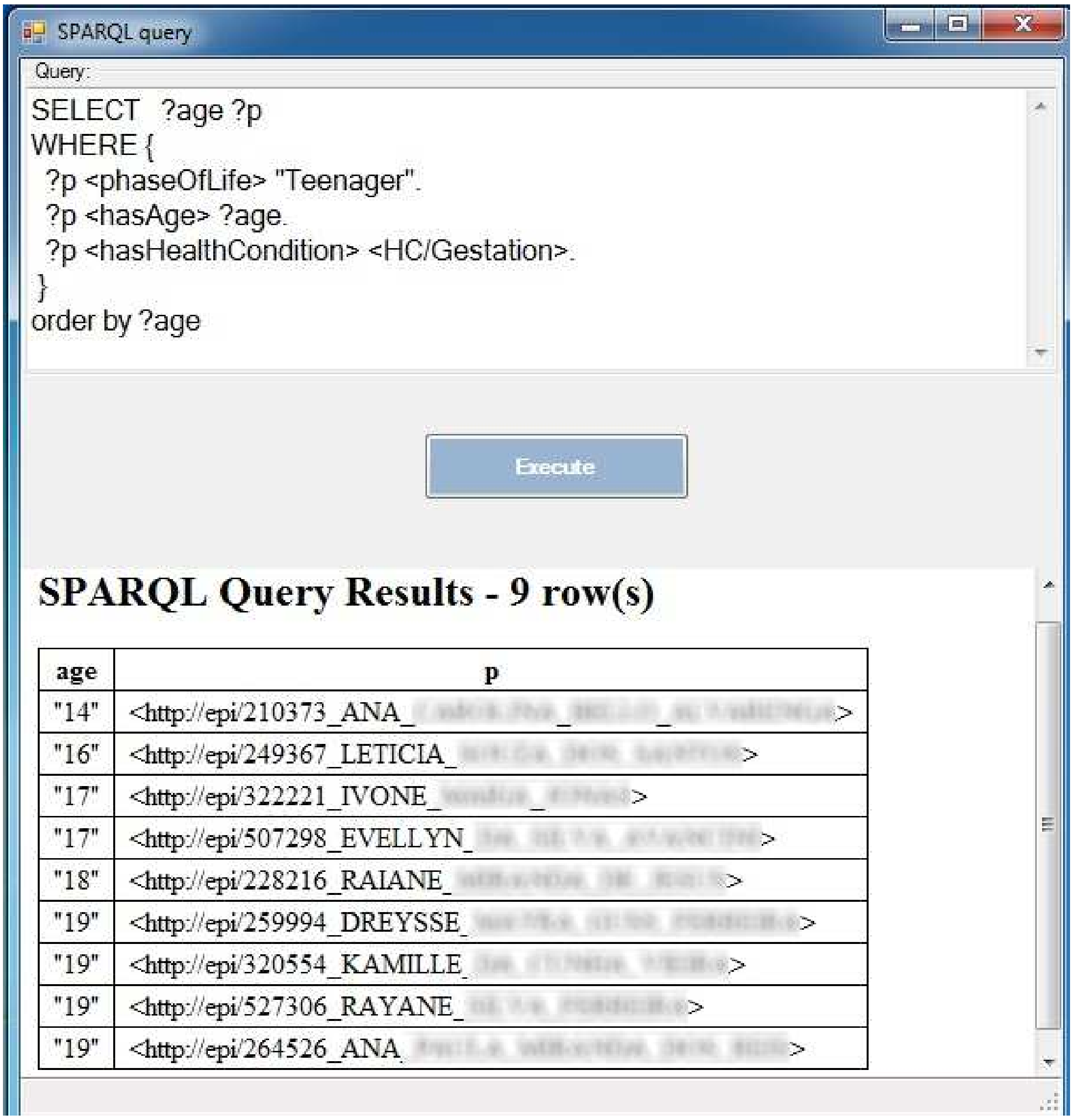


Figure 5.14: SPARQL to answer the competency question - ***CQ 3***.

5.5.5 Answering the competency question - *CQ 5*

The competency question ***CQ 5*** (What are the patient has no hypertension, but has the epidemiologic profile (eg, age, sex) of a patient with hypertension?), aims to identify patients with epidemiological profile of hypertension.

For this our example, we are identifying these patients through the meta-rules created and presented in Figure 5.8 (page 83). In Figure 5.16 is an example of SPARQL query to identify all patients who have epidemiological profile of the health condition hypertension.

In this example we are using the predicate *<hasEpidemiologicalProfile>*, with the object *<HC/Hypertension>*, so that the meta-rules created to be used by the system. As a result we have 259 patients who have the epidemiological profile of patients with hypertension, ie, they

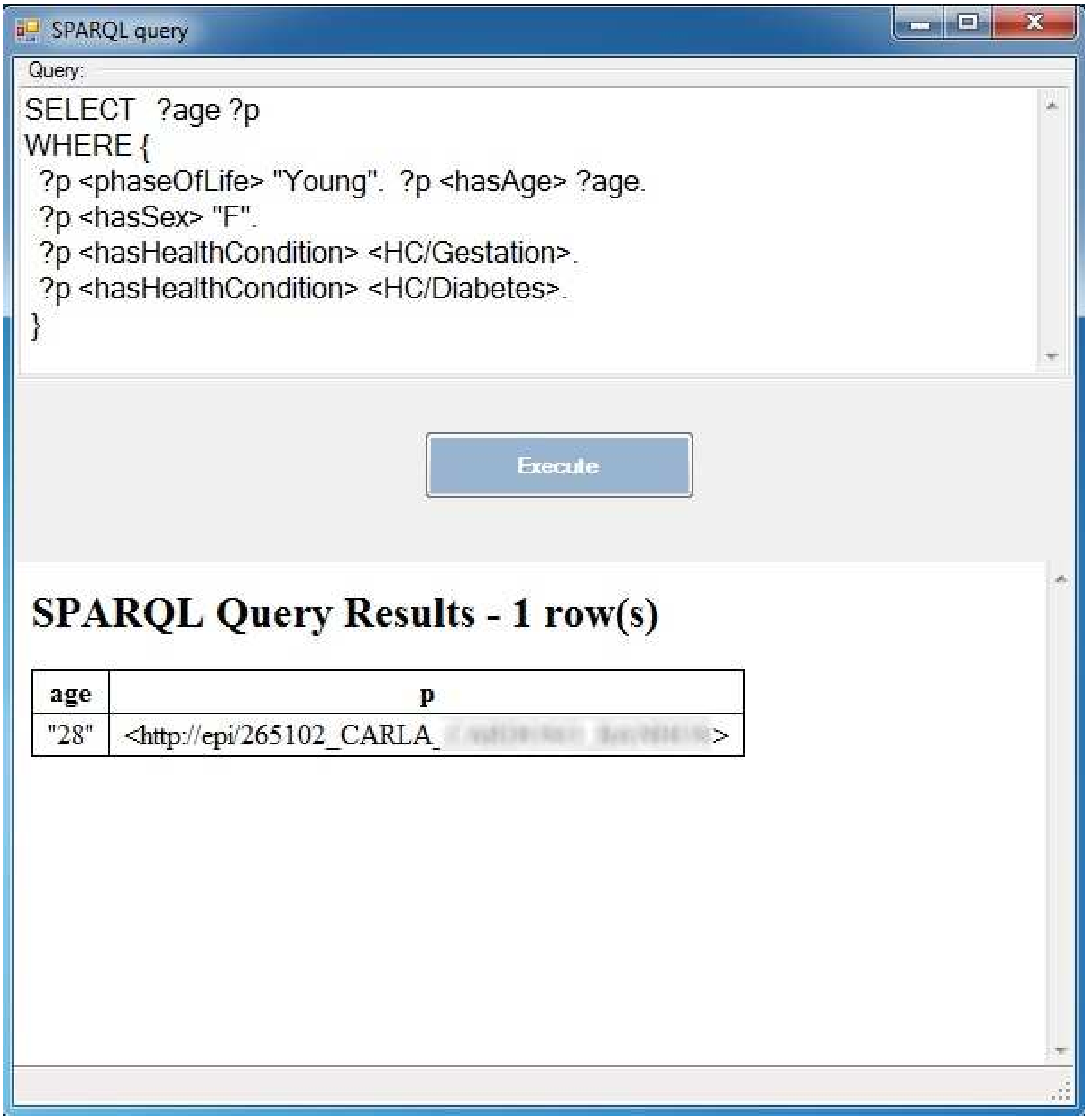


Figure 5.15: SPARQL to answer the competency question - ***CQ 4***.

are patients who do not have the health condition of hypertension today, but has the probability (trend) to have this health condition in the future.

In order to better understand the result, we detail a random case of a record returned by the query, in our case the patient's record was Lucia, the sixth record returned. Consulted this patient: age, their health conditions and the risk of your family.

The SPARQL query is shown in, Figure 5.17. The result was that the patient is age of 25 years, has no associated health condition and her family has medium risk.

These results confirm that the meta-rule was properly verified by the system, being able to identify meta-rule in which the record has been validated, it was the second meta-rule in Figure 5.8 (page 83), which treats the case of families of medium risk.

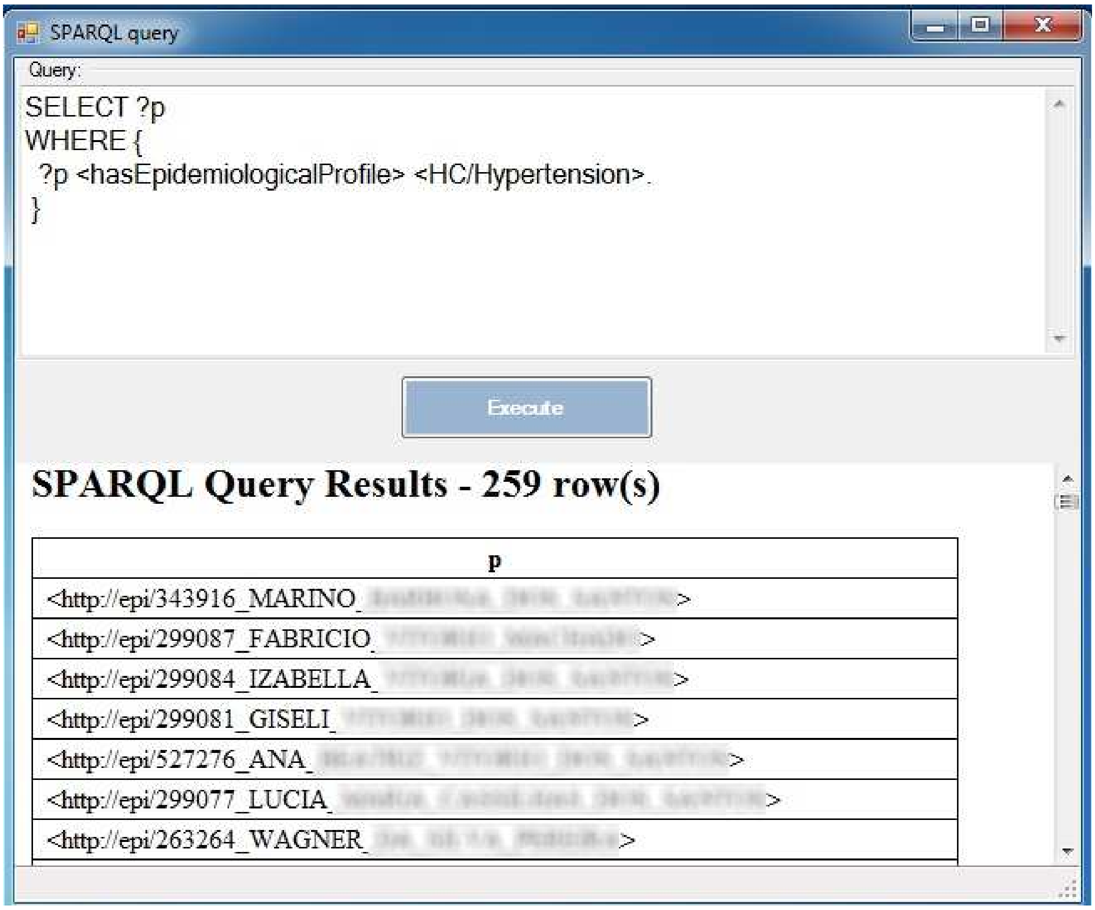


Figure 5.16: SPARQL to answer the competency question - ***CQ 5***.

5.5.6 Answering the competency question - *CQ 6*

Through the ***CQ 6*** (What are the elderly who live alone?) we identify people who are elderly, and are vulnerable because they live alone, regardless of the health condition.

In Figure 5.18, is shown SPARQL query to answer this question. First we obtained members of all families (*?f <hasMembersOfFamily> ?members*), then filtered by the members who are elderly (*?members <phaseOfLife> "Elderly"*), then add one more filter to only consider people who live alone, in other words, which has the family members just one person (*?f <hasNumbersOfMembers> ?n. FILTER regex(?n, "^1$")*).

As a result 473 records are displayed. This example shows people sex, male or female. If the intention was to query only elderly female living alone, we need only to add the gender restriction in the query: *?members <hasSex> "F"*.

As the above example, we could perform other queries like: Elderly who live alone and have any health condition by adding the following RDF triple, *?members <hasHealthCondition> <HC/Hypertension>*, for this example, we are considering the health condition of hypertension.

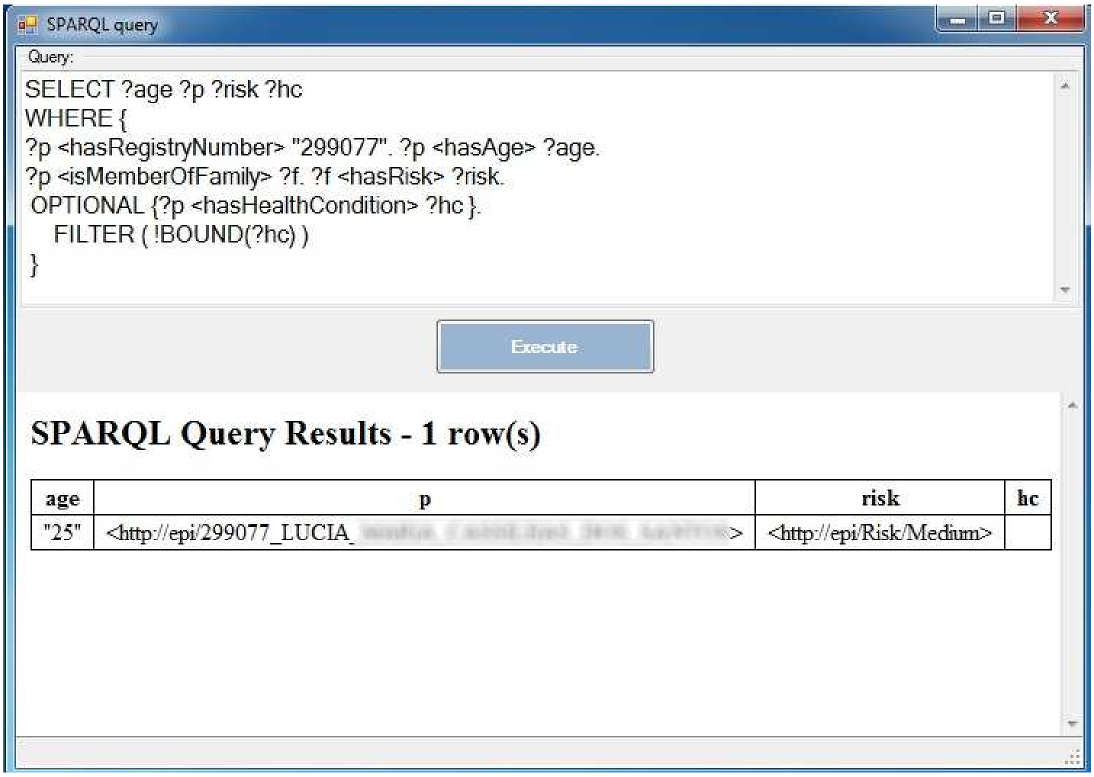


Figure 5.17: Detail of the result obtained by the SPARQL query the competency question - ***CQ 5***.

5.5.7 Answering the competency question - *CQ 7*

Through the ***CQ 7*** (What are the families who did not receive visits in the month?), we were able to monitor families that are not being accompanied with monthly home visits.

Figure 5.19, shows a SPARQL query to answer this question. For this example, we recover all the families of micro area 7 (*?f <hasMicroArea> "7"*), which did not receive visits (*?f <hasNoVisit> ?v*) in the month February (*?v <hasMonth> "2"*) of 2012 (*?v <hasYear> "2012"*).

As result, we have that 156 families in the micro area 7 did not receive visits in February of 2012. This information is important for the scheduling of visits by Community Health Agent that is performed by Nurse. Knowing the families that were not accompanied in the month, the Nurse can prioritize for that home visits to be held next month.

Besides knowing all the families who did not receive home visits in the month, it is very important to know if any family of high risk has not been visited. To realize this filter, we may simply add the following RDF triple: *?f <hasRisk> <Risk/High>*.

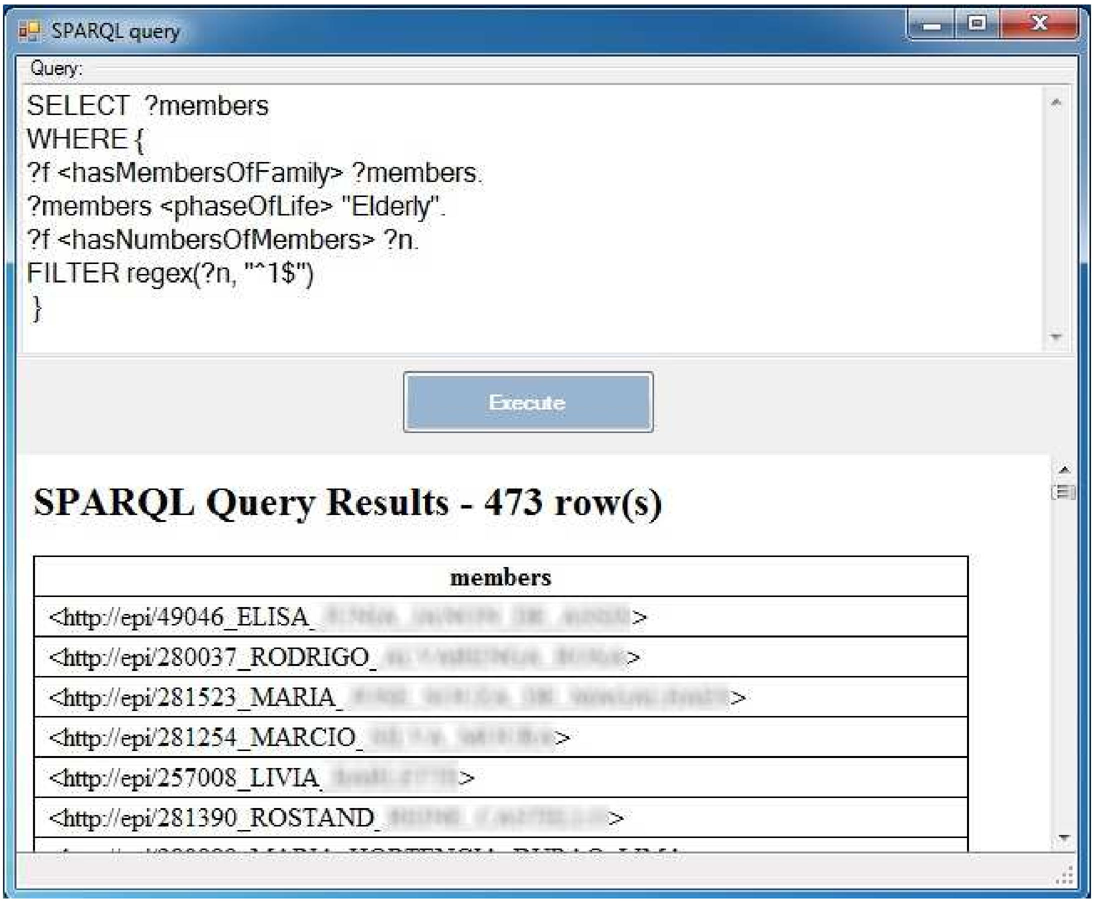


Figure 5.18: SPARQL to answer the competency question - ***CQ 6***.

5.6 Related Work

In SHI et al.(2011) is shown the ScienceWeb. The objective of the system is to provide answers to qualitative queries that represent the evolving consensus of the community of researchers. The system allows qualitative queries such as: "Who are the groundbreaking researchers in data mining?"or "Is my record tenurable at my university?".

In ScienceWeb, the knowledge base is represented using the web ontology language (OWL) and the ontology reasoning system is performed with support custom rules. Customized rules are used when qualitative descriptors involve transitive reasoning as in "What is the PHD advisor genealogy of professor x?"or complex queries that involve algebraic computation across a populations such as in "groundbreaking", the custom rules are written in SWRL (Semantic Web Rule Language) .

The similarity with our work is that this system also provides answers to queries that are submitted to the system. In our study, questions are related to the epidemiological profile, and the knowledge of the ontology is represented in RDF and additional rules are defined in Euler.

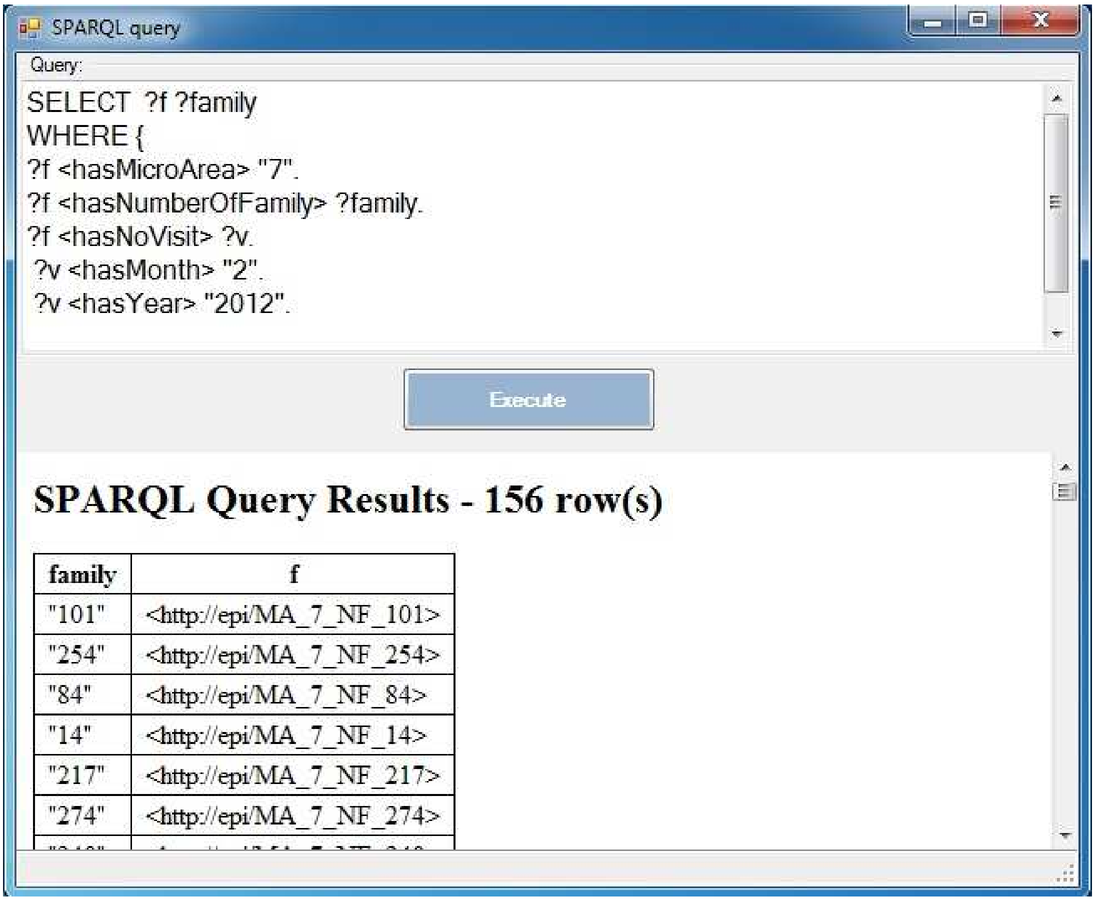


Figure 5.19: SPARQL to answer the competency question - ***CQ 7***.

In comparison with the work by SHI et al., the custom rules are similar to the rules presented in this work. In the work by SHI et al., it is not shown any proposal similar to the meta-rules where rules are created from machine learning algorithms, as in our work.

ZHANG; SILVESCU; HONAVAR(2002) present the ontology-driven decision tree learning algorithm to learn classification rules at multiple levels of abstraction. There are applications where users often need to examine data in different contexts from different perspectives and at different levels of abstraction, there is no single universal ontology that can serve all users, or for that matter, even a single user, in every context. Hence, methods for learning from ontologies and data are needed to support knowledge acquisition from heterogeneous distributed data.

Most learning algorithms for data-driven induction of pattern classifiers (e.g., the decision tree algorithm), typically represent input patterns at a single level of abstraction, usually in the form of an ordered tuple of attribute values. They typically assume that each pattern belongs to one of a set of disjoint classes. Thus, any relationships that might exist between classes (e.g., a hierarchically nested class structure) are ignored. However, in many applications of inductive learning, for example, scientific discovery, users often need to explore a data set at multiple levels of abstraction, and from different points of view. Each point of view corresponds to a set of ontological (and representational) commitments regarding the domain of interest. Thus, there is a need for inductive algorithms that can exploit explicit ontologies to examine at different levels of abstraction.

In contrast to this dissertation, in ZHANG; SILVESCU; HONAVAR the learning, which is the identification of rules obtained by decision tree, is aided by the use of one or more ontologies, allowing relationships to be established between the terms of the ontologies.

For example, to identify the shopping profile of some customers, two ontologies are used (Beverages and Snacks). Using the decision tree algorithm in the data set representing the purchases of these customers it is identified the type of beverage or snack that customers tend to consume according to age. Through the relationship between ontologies, it was identified that consumers who bought Cereal (sub-type of snack), also bought Skim Milk or Low Milk (sub-type of milk, which is sub-type dairy beverage), thus the rule that customers who buy Cereal, also buy dairy beverage can be defined.

5.7 Conclusions

In this chapter we present the system developed to allow performing queries on the computational model created using RDF to obtain information about epidemiological profile.

To give answers the competency questions (see Chapter 4, Table 4.3), the system uses the concepts of the ontology using the SPARQL query language, to provide the answer. Furthermore, for the system to be able to answer epidemiological questions it is important to be able to assess health conditions, so on, in which patients are inserted. With this objective, this study used a decision tree.

Decision trees were used to classify how you define the epidemiological profile of a given population, based on characteristics such as: age, sex, health condition, so on.

For this analysis, we use the algorithm called C4.5 decision tree (QUINLAN, 1993). In our study, we used the results of decision tree with pruning, that reduces the size of decision trees by removing sections of the tree that provide little power to classify instances.

Besides the use concepts of the ontology, other concepts have been defined using rules and meta-rules, allowing the execution of inference and detection of inconsistencies in the ontology. In our study the difference between rule and meta-rule is that: Rules are the concepts included in the model which were not obtained in an automated way using the decision tree. Meta-rules are concepts added to the model from the results obtained by the decision tree.

It is very important to highlight that the study of decision tree, can not be taken as an absolute truth. For this reason, we call meta-rules, not rules, the meta-rules can change over time, for example, a meta-rule that describes a person with the possibility of having diabetes can change the measure that epidemiological profile of the territory changes. When this occurs the meta-rule must be redone.

Through this study we seek to offer a set of tools that can support the study, analysis and understanding of the epidemiological profile of a given territory of health.

The next chapter presents the conclusions of the work, showing its contributions, positive and negative points, limitations, difficulties in their development as well as future work.

6 Conclusion

This chapter presents the conclusions of the work, showing their contributions. Finally, we present limitations and prospects for future work that may arise based on this work.

6.1 Results of the Dissertation

Epidemiology has become important for primary health care, providing knowledge of the phenomena of health and disease, focusing not only on medical intervention(VIANA; POZ, 2005). Epidemiology is used for calculations and statistical techniques, in the studies the distribution and determinants of health problems in human populations.

Characterized as an area strongly based on knowledge, medicine has been greatly helped by systems of knowledge management. Among these systems, we highlight those aimed at diagnosis, health promotion and decision support, the planning of medical therapy and classification (BOSE, 2003).

The large amount of works involving the use of Ontologies, mainly motivated by the need for semantic interoperability among domain models and applications seeks to take advantage of ontology-based system, that are: a) reuse knowledge instead of reusing software; b) reuse and share application domain knowledge using a common vocabulary across heterogeneous software platforms and programming languages; and c) focus on domain structure, preventing them from being distracted by implementation details (GUIZZARDI et al., 2009).

The ontology development begins with the definition of a set of questions, named competency questions. Competency questions is a way to determine the scope of the ontology and outline a list of questions that the ontology will be able to respond. Similarly, requirements engineering aims at the definition of the behavior of the software system, considering functional and non-functional requirements (GRÜNINGER; FOX, 1995).

Several studies have been developed in the literature for the development of ontologies and their approaches to specify competency questions, we can identify that these methodologies still do not satisfactorily consider guidelines for the identification of competency questions. Consequently, the ontology engineer remains so far with very limited resources to capture the requirements of the ontology in the process of ontology engineering.

This dissertation discussed these aspects and our main contribution is a system to study the epidemiological profile in a basic health care unit, being that this system applies an ontology as basis for modeling and querying the epidemiological information.

In order to develop the system, a goal-oriented methodology based on Tropos is applied to capture and model the competency questions (i.e. the requirements) of an ontology. As far as we know, this is the first attempt to provide a visual solution to support these early stages of ontology engineering, other contributions are:

- A case study developed in a real health care setting, which illustrates and provides some interesting insights on the use of the methodology in practice;
- The use of decision tree to support learning in an ontology-based system to define the epidemiological profile;
- Development of specific functionality to support the study, analysis and understanding of the epidemiological profile of a given health territory.

The approach presented here can also be applied to other types of system, whether in or outside the medical field. For example, we could use this approach to e-commerce system, where through the competency questions defined by using Tropos. As an example of a competency question could be the purchasing profile of a customer's web site.

6.2 Limitations and Future Work

The preliminary use of the methodology proposed in this dissertation, through case study allowed the identification of several opportunities for improving aspects that still have limitations, in addition to motivating the creation of new features.

The main points glimpsed due to some limitations already identified, are:

- Provide a methodology which considers both the case in which the ontology is automated and the case in which it serves as a conceptual model to mediate and support people's communication and cooperation, without the need of a system. In this dissertation, we focused in the former, i.e. we model an ontology-based system;

- Ontology reuse is one of the steps we intend to add in the methodology, allowing to reuse existing ontologies instead of building one from scratch.
- For the case study, the use of *SemWeb.NET* with a large number of triple RDF during queries SPARQL was extremely slow, so we intend to evaluate other approaches, such as the Virtuoso (VIRTUOSO, 2012) tool.

As future work, aiming at continuing this line of work, we propose:

- Extension of the system to use other data stored in the LOD (Linked Open Data) cloud, that contribute to assessment of the epidemiological profile. The city hall and the department of health of our city is interested in using this system to support their practical work related to the Family Health Program.
- Develop a validation experiment in which we hope to compare our methodology with some existing approaches, by having real modelers in a practical assignment.

References

ANDRADE, S. M.; SOARES, D. A.; JUNIOR, L. C. *Bases da Saúde Coletiva*. [S.l.]: Londrina: UEL, 2001.

BARRETO, M. L. Papel da epidemiologia no desenvolvimento do sistema Único de saúde no brasil: histórico, fundamentos e perspectivas. *Revista Brasileira de Epidemiologia*, v. 5, 2002.

BEAGLEHOLE, R.; BONITA, R.; KJELLSTROM, T. *Epidemiologia Básica*. 2. ed. [S.l.]: São Paulo: Santos, 2003.

BERNARAS, A.; LARESGOITI, I.; CORERA, J. Building and reusing ontologies for electrical network applications. *The European Conference on Artificial Inteligence, ECAI-96*, p. 298–302, Mar. 1996.

BIZER, C.; CYGANIAK, R.; GAUß, T. The rdf book mashup: From web apis to a web of data. *Proceedings of the 3rd Workshop on Scripting for the Semantic Web (SFSW2007)*, 2007.

BOEHM, B. Industrial software metrics: top ten list. IEEE, Software, Sept 1987.

BORNSTEIN, V. J.; MARIA, H.; LEAL, S. *Agentes comunitários de saúde : A reconstrução do conceito de risco no nível local*. [S.l.]: Interface - Comuniação, Saude, Educação, 2010. 93-101 p.

BOSE, R. Knowledgemanagement-enabled health care managementsystems: capabilities, infrastructure, and decision-support. *Expert Systems with Applications*, v. 24, p. 59–71, January 2003.

BOTTI, M. L. O aprender organizacional : relato de experiência em uma unidade básica de saúde. *Saúde e Sociedade*, v. 15, p. 107–114, 2006.

BRASIL. Ministério da saúde. portaria ministerial n^{o}. 648, de 28 de março de 2006. aprova a política nacional de atenção básica. *Pacto pela Saúde, Ministério da Saúde. Brasília*, v. 4, 2006.

BRESCIANI, P. et al. Tropos: An agent-oriented software development methodology. *International Journal of Autonomous Agents and Multi Agent Systems*, v. 8, p. 203–236, 2004.

BROWN, J.; DUGUID, P. The social life of information. *Harvard Business School Press, Cambridge MA*, 2000.

BRUSA, G.; CALIUSCO, M.; CHIOTTI, O. A process for building a domain ontology: an experience in developing a government budgetary ontology. Proc. Second Australasian workshop on Advances in ontologies (AOW '06), Australian Computer Society, v. 72, p. 7–15, 2006.

CAMPOS, M. L. A. *O papel das definições na pesquisa em ontologia*. 1. ed. [S.l.]: Perspect. ciênc. inf., Belo Horizonte, 2010.

CANDAN, K.; LIU, H.; SUVARNA, R. Resource description framework: metadata and its applications. *SIGKDD Explor. Newsl.*, ACM, New York, NY, USA, v. 3, n. 1, p. 6–19, jul. 2001. ISSN 1931-0145. Disponível em: <http://doi.acm.org/10.1145/507533.507536>.

CARROL, J. Scenario-based design: Envisioning work and technology. In System Development by John M. Carroll, 1995.

CARVALHO, B. G. *Organização do sistema de saúde no Brasil*. [S.l.]: Londrina: UEL, 2001.

CASTRO, J.; KOLP, M.; MYLOPOULOS, J. Towards requirements-driven information systems engineering: the tropos project. *Information Systems, Amsterdam, The Netherlands*, v. 27, n. 6, p. 365–389, 2002.

COELHO, F. L. G.; SAVASSI, L. C. M. Aplicação da escala de risco familiar como instrumento de priorização das visitas domiciliares. Revista Brasileira de Medicina de Família e Comunidade, v. 1, p. 19–26, 2004.

CRANEFIELD, S.; PURVIS, M. Uml as an ontology modelling language. In: *Proceedings of the IJCAI-99, Workshop on Intelligent Information, 16th International Joint Conference on AI*. Stockholm, Sweden, Sweden: [s.n.], 1999.

CRUZ, S. *Avaliação do Processo de Implantação da Estratégia de Saúde da Família: O Caso de Resistência*. Dissertação (Mestrado) — UFES/ES, 2007.

DAVENPORT, T.; JARVENPAA, S.; BEERS, M. Improving knowledge work processes. *Sloan Management Review*, v. 37(4), p. 53–65, 1996.

DAVIS, A. M. *Software Requirements*. [S.l.]: Prentice Hall, Englewood Cliffs, New Jersey, 1993.

FALBO, R. Experiences in using a method for building domain ontologies. In: *Proceedings of the 16th International Conference on Software Engineering and Knowledge Engineering, International Workshop on Ontology In Action*. Banff, Canada, Canada: [s.n.], 2004.

FILHO, N. A.; ROUQUAYROL, M. Z. *Introdução à Epidemiologia*. [S.l.]: Rio de Janeiro: MEDSI, 2002.

FOX, M. The tove project towards a common-sense model of the enterprise, industrial and engineering applications of artificial intelligence and expert systems. Springer Berlin / Heidelberg, p. 25–34, 1992.

FRANCO, T.; MERHY, E. Psf: Contradições e novos desafios. *Conferência Nacional de Saúde On-Line*, 1999.

GASEVIC, D.; DJURIC, D.; DEVEDZIC, V. Model driven engineering and ontology development. Springer, Berlin, 2009.

GIORGINI, P. et al. The tropos metamodel and its use. *"Informatical journal"*, 2005.

GÓMES-PÉREZ, A.; FERNÁNDES-LÓPEZ, M.; CORCHO, O. Ontological engineering. Springer-Verlag, New York, 2004.

GÓMEZ-PÉREZ, A.; SUÁREZ-FIGUEROA, M. C. Scenarios for building ontology networks within the neon methodology. *Proceedings of the fifth international conference on Knowledge capture (K-CAP '09). ACM, New York, NY, USA*, p. 183–184, 2009.

GRÜNINGER, M.; FOX, M. Methodology for the design and evaluation of ontologies. Proc. IJCAI-95 Workshop on Basic Ontological Issues in Knowledge Sharing, Montreal, August 19-20th 1995.

GRUBER, T. R. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 1993.

GUARINO, N. Formal ontology in information systems. In: *Proceedings of the 1st International Conference June 6-8, Trento, Italy*. [S.l.: s.n.], 1998.

GUIZZARDI, G. On ontology, ontologies, conceptualizations, modeling languages, and (meta)models. In: *Proceedings of the 2007 conference on Databases and Information Systems IV: Selected Papers from the Seventh International Baltic Conference DB&IS'2006*. Amsterdam, The Netherlands, The Netherlands: IOS Press, 2007. p. 18–39. ISBN 978-1-58603-715-4.

GUIZZARDI, G. et al. On the importance of truly ontological distinctions for ontology representation languages: An industrial case study in the domain of oil and gas. In Proceedings of the 14th International Conference on Exploring Modeling Methods in Systems Analysis and Design (EMMSAD´09), Amsterdam, The Netherlands, Lecture Notes in Business Information Processing (LNBIP), 2009.

GUIZZARDI, R. Ontological foundations for structural conceptual models. *Ph.D. Dissertation - University of Twente, The Netherlands*, 2005.

GUIZZARDI, R.; LUDERMIR, P.; SONA, D. A recommender agent to support knowledge sharing in virtual enterprises, in protogeros. Agent and Web Service Technologies in Virtual Enterprises, Idea Group Publishing, 2007.

GUIZZARDI, R. S. S.; PERINI, A. Analyzing requirements of knowledge management systems with the support of agent organizations. *Journal of the Brazilian Computer Society (JCBS). Special Issue on Agents Organizations*, v. 11, p. 51–62, July 2005.

HAYES, P. The naive physics manifesto. *In D. Ritchie (Ed.) Expert Systems in Microeletronics age*, Edinburgh University Press, p. 242–270, 1978.

HAYES, P. Naive physics i: Ontology for liguids. Hobbs & Moore, p. 71–107, 1985.

HOLLAND, P.; LEINHARDT, S. Perspectives on social network research. New York: Academic Press, 1979.

JUNIOR, M. D. *Epidemiologia nos Municípios: Muito além das normas*. [S.l.]: São Paulo: HUCITEC, 2003.

JURETA, I.; FAULKNER, S.; SCHOBBENS, P. A more expressive softgoal conceptualization for quality requirements analysis. Arizona, 2006.

KAVAKLI, E.; LOUCOPOULOS, P. Goal driven requirements engineering: Evaluation of current methods. 2003.

KLYNE, G.; CARROLL, J. J. *Resource Description Framework (RDF): Concepts and Abstract Syntax*. February 2004. Http://www.w3.org/TR/rdf-concepts/. Accessed: 05/05/2012.

LAMSWEERDE, A. Goal-oriented requirements engineering: A guided tour. 2001.

LENAT, D. B. Cyc: a large-scale investment in knowledge infrastructure. Commun. ACM, v. 38, p. 33–38, Nov. 1995.

LENAT, D. B. et al. Cyc: toward programs with common sense. Commun. ACM, v. 33, p. 30–49, Aug. 1990.

LIASKOS, S. et al. Integrating preferences into goal models for requirements engineering. Proc. 10th International Requirements Engineering Conference (RE-10), 2010.

LINKED-DATA. *Linked Data - Connect Distributed Data across the Web*. abr 2012. Http://linkeddata.org/. Accessed: 29/04/2012.

LUTZ, R. R. Analyzing software requirements errors in safety-critical embedded systems. Proc. ACM SIGSOFT Symposium on the Foundations of Software Engineering, New York, NY, p. 126–133, Dec 1993.

MINISTÉRIO DA SAÚDE. *O SUS de A a Z: Garantindo Saúde nos Municípios*. 3. ed. [S.l.]: Brasília-DF:MS, 2009.

MYLOPOULOS, J.; CHUNG, L.; NIXON, B. Representing and using non-functional requirements: a processoriented approach. IEE Transactions on Software Engineering, v. 18, n. 6, June 1992.

NEIGER, D.; CHURILOV, L. Goal-oriented business process engineering revisited: a unifying perspective. 2004.

NOY, F. N.; GUINNESS, D. L. Ontology development 101: a guide to create your first ontology. *Stanford Knowledge Systems Laboratory Technical Report KSL-01-05*, 2001.

PAIM, J. *Saúde, Política e Reforma Sanitária*. [S.l.]: Salvador: Institudo de Saúde Coletiva, 2002.

QUILITZ, B.; LESER, U. Querying distributed rdf data sources with sparql. *Proceedings of the 5th European Semantic Web Conference (ESWC2008)*, 2008.

QUINLAN, J. R. *C4.5: Programs for Machine Learning*. [S.l.]: Morgan Kaufmann Publishers, 1993.

RUSSEL, S.; NORVIG, P. *Artificial Intelligence, a modern Approach*. 2. ed. [S.l.]: Prentice Hall, 2003.

SHI, H. et al. Comparison of ontology reasoning systems using custom rules. In: *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*. New York, NY, USA: ACM, 2011. (WIMS '11), p. 16:1–16:9. ISBN 978-1-4503-0148-0.

SILVA, A. S.; LAPREGA, M. R. *Avaliação crítica do Sistema de Informação da Atenção Básica (SIAB) e de sua implantação na região de Ribeirão Preto, São Paulo, Brasil*. [S.l.]: Cadernos de Saúde Pública. Rio de Janeiro, 2005. 1821-1828 p.

SOARES, A. Towards ontology-driven information systems: Guidelines to the creation of new methodologies to build ontologies. *Ph.D. Dissertation - The Pennsylvania State University*, 2009.

SPARQL-XML. *SPARQL Query Results XML Format*. May 2012. Http://www.w3.org/TR/rdf-sparql-XMLres/. Accessed: 05/05/2012.

STEFANELLI, M. Knowledge and process management in health care organizations. *Methods of Information in Medicine*, v. 43(5), p. 525–535, 2004.

SUÁREZ-FIGUEROA, M. C. et al. Neon methodology for building contextualized ontology networks. *NeOn Deliverable D5.4.1.*, NeOn Project, February 2008.

SUS. *Seleção das doenças de notificação compulsória: critérios e recomendações para as três esferas de governo*. [S.l.]: Brasília, Ministério da Saúde, 1998.

TAUBERER, J. *Semantic Web/RDF Library for C#.NET*. abr 2012. Http://razor.occams.info/code/semweb/. Accessed: 29/04/2012.

USCHOLD, M.; KING, M. Towards a methodology for building ontologies. *Proc. IJCAI-95 Workshop on Basic Ontological Issues in Knowledge Sharing*, July 1995.

USCHOLD, M. et al. The enterprise ontology. *Knowl. Eng.*, p. 31–89, Mar. 1998.

VIANA, A. L.; POZ, M. R. D. A reforma do sistema de saúde no brasil e o programa de saúde da família. *Physis: Revista de Saúde Coletiva*, v. 15, p. 225–264, 2005.

VIRTUOSO. *Virtuoso Universal Server*. May 2012. Http://virtuoso.openlinksw.com/. Accessed: 01/05/2012.

YAMAMOTO, S. et al. Goal oriented requirements engineering: Trends and issues. *IEICE - Transactions on Information Systems E89-D*, 2006.

YU, E. S. Modelling strategic relationships for process reengineering. *Ph.D. Dissertation - University of Toronto, Toronto*, 1995.

ZHANG, J.; SILVESCU, A.; HONAVAR, V. G. Ontology-driven induction of decision trees at multiple levels of abstraction. In: *Proceedings of the 5th International Symposium on Abstraction, Reformulation and Approximation*. London, UK, UK: Springer-Verlag, 2002. p. 316–323.