BUILDING A FAMILY ONTOLOGY TO MEET CONSISTENCY CRITERIA

TAN MEE TING

A thesis submitted in partial
fulfillment of the requirement for the award of the
Degree of Master of Computer Science (Web Technology)

Faculty of Computer Science and Information Technology
Universiti Tun Hussein Onn Malaysia

FEBRUARY 2015

ABSTRACT

Semantic web is an extension of the current web in which the existing information on the web are organized and encoded more meaningfully using ontology language, thus enabling effective communication among machines and humans. Ontology is the backbone of the semantic web that contributes to knowledge sharing among intended parties over distributed systems around the world. In the past few years, semantic web has been widely accepted by a variety of fields for better knowledge representation, communication, sharing and reasoning on the web. Now, there are existing genealogical ontologies proposed by different groups of researchers once semantic web has emerged as third generation of the web. However, existing ontologies still lack certain important concepts and properties to support the domain of family relations. This may lead to the inability of the ontology to deliver full potential of exchanging family history information among all interested parties. Moreover, existing ontologies do not employ the full potential of SWRL rules to reason the individuals within the ontology. The main aim of this research is to build a new Family Ontology which obeys the consistency criteria. Consistency checking ensures there are no contradictory concepts found within the resulting ontology. The consistency of Family Ontology will be evaluated using FACT++, HermiT and Pellet reasoners. By augmenting the additional axioms and testing the resulting ontology thoroughly using reasoner tools, the proposed Family Ontology is expected to achieve a consistency of 100%.This research is meaningful and significant to all humans since everyone has his or her own unique family history. The proposed ontology also facilitates effective and efficient communication among all intended parties since shared vocabularies and standards are employed by the proposed ontology.

ABSTRAK

Web Semantik ialah teknik terbaru yang membolehkan data pada Web zaman terkini disusun dan diaturcara secara bermakna dengan menggunakan bahasa ontologi. Ontologi struktur memudahkan komunikasi berlangsung secara efektif antara komputer dan manusia. Ontologi umpama tulang belakang bagi Web Semantik yang menyumbang kepada perkongsian maklumat antara pihak-pihak tertentu melalui rangkaian Internet di seluruh dunia. Web Semantik telah mendapat sambutan meluas dalam pelbagai bidang pada hari ini dan ia merupakan cara terbaik untuk mengekodkan data-data bagi tujuan komunikasi, perkongsian dan reasoning pada Web. Terdapat beberapa genealogi ontologi telah dicipta sejak kebelakangan ini dan kesemuanya telah dicadangkan oleh penyelidik-penyelidik berlainan apabila Web Semantik muncul sebagai Web generasi ketiga. Namun, ontologi yang sedia ada masih kekurangan konsep dan relasi penting bagi menyokong keluarga domain. Hal ini menyebabkan ontologi tidak mampu menunjukkan potensi sepenuhnya dalam perkongsian maklumat sejarah keluarga antara semua pihak. Tambahan pula, genealogi ontologi yang sedia ada tidak menggunakan fungsi peraturan SWRL sepenuhnya bagi tujuan reasoning pada individu-individu dalam ontologi. Matlamat utama kajian ini adalah untuk menghasilkan satu Ontologi Keluarga yang memenuhi kriteria konsisten. Ujian konsisten memastikan tiada konsep yang bertentangan di dalam ontologi. Konsistensi akan dinilai dengan menggunakan FACT++, HermiT and Pellet. Dengan memasukkan aksioma tambahan dan memeriksa ontologi secara teliti, Ontologi Keluarga yang dicadangkan dianggap telah mencapai konsistensi 100 peratus. Kajian ini amat bermakna dan agak penting terhadap semua manusia kerana setiap orang memiliki sejarah keluarga mereka yang unik. Ontologi yang dicadangkan ini turut membolehkan komunikasi berlangsung secara berkesan dan efektif antara semua pihak kerana kosa kata dan standard yang sama sentiasa dirujuk oleh semua pihak.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

ABOX	-	Assertional Box
CBR	-	Case Based Reasoning
CDSS	-	Clinical Decision Support System
DL	-	Description Logic
EL	-	Expressive Language
FACT++	-	Fact Plus Plus
GEDCOM	-	Genealogical Data Communication
GENTECH	-	Genealogical Data Model
GNU	-	General Public License
HTML	-	Hypertext Markup Language
HTTP	-	Hyper Text Transfer Protocol
KACTUS	-	Knowledge About Complex Technical Systems for Multiple Use
LGPL	-	Lesser General Public License
OWL	-	Web Ontology Language
RACER	-	Renamed Abox and Concept Expression Reasoner
RBR	-	Rule Based Reasoning
RDF	-	Resource Description Framework
RDFS	-	Resource Description Framework schema
SNOMED	-	Systematized Nomenclature of Human Medicine
SPARQL	-	Simple Protocol and Resource Description Query Language
SQL	-	Structured Query Language
SWRL	-	Semantic Web Rule Language
TBOX	-	Terminological Box
UNA	-	Unique Name Assumption
W3C	-	World Wide Web Consortium

CHAPTER 1

INTRODUCTION

1.1 Research Background

In genetic context, a family is often regarded as a group of people who have blood relations with each other or a group of descendents from a common ancestor. Basically, a unit of family is described as living together in one household. Apart from residing in a shared physical location, they usually share many other common elements in general which include ancestors, traditions, religions, lifestyles, environments and even genes that contribute to the risk of hereditary diseases. Typically, for viewing and readability purposes, the family relationships for a unit of family over few generations can be visualized using a family tree. A family tree is a chart normally used for representing the family relations in a conventional tree structure with interconnected nodes linked together via family relations. Family history information can be utilized for various purposes. Apart from being used to trace the ancestors of a person, a doctor can also use this particular information to predict family health problems since family relations are the common factors for most of the hereditary diseases. For instance, a completed genealogical chart can be exploited or extended to support multiple kinds of functions in medical or social work. In the medical field, this goal can be achieved by annotating additional data such as medical conditions of family members who have suffered from certain diseases. By having a precise parental health history, a doctor is able to identify the risk of a person developing certain diseases at an early stage and take necessary precautions earlier to avoid and minimize the risk of those diseases [1, 2, 3].

The risk of a disease being transmitted by parents to their children becomes higher when many of the family members were affected by certain common diseases. If the family members involved are first or second degree relatives and the diseases were developed at young age, then the probability of a child inheriting the same disease as their parents will increase further [1, 2]. The importance of family medical history has long been recognized in caring a patient [4]. By examining the family medical history, a doctor is able to make quick and effective decisions on immediate actions which should be taken to minimize the risk of particular diseases. In addition, family medical history data can assist a doctor in identifying family members who have higher risk of developing certain disease, deciding whether the family members should obtain a specific genetic test, determining the type and frequency of screening tests and assess the risk of passing those diseases to their children.

However, before having a completed family medical history, the first step will be building a precise and consistent genealogical chart or family history. There were aggressive researches in recent years on genealogical ontology after the semantic web has emerged as third generation of the web but some improvements can still be made towards the existing works. Improvements can be made towards the consistency, reusability, taxonomy and inference of existing family ontologies. In semantic web, ontology is used to encode the knowledge on the web in a semantic manner. According to Gruber [5], ontology is a formal, explicit specification of a shared conceptualization. This also means that ontology codifies relevant concepts of one phenomenon into machine readable format where the encoded knowledge is understood and agreed upon by large communities in general. Moreover, recent research has found out that ontology is the most powerful tool to represent knowledge formally [6, 7].This fact is proven when there were considerable numbers of domain experts who initiated their attempts to employ ontology as their representation languages in both medical and genealogical related applications. Applications under genealogical field were clinical knowledge-based systems such as SNOMED [8], Gene Ontology [9] and National Cancer Institute Thesaurus[10].

Those initiatives have shown that the value of ontology is gradually being recognized by the public. In fact, ontology is not merely accepted widely in genealogical

and medical areas but in reality it has also been adapted in a variety of fields. For now, ontology has even become the alternative way for search engines, e-commerce web sites, WorldNet, artificial intelligence and multi-agent systems. Actually, there are multiple factors which contribute to progressive researches on ontologies and creation of ontologies for various domains. Encoding pieces of knowledge using ontology is advantageous since ontology is capable of sharing common understanding of information among different parties in a community, research group, organization and software agents across the internet. Variety of standards and heterogeneous data employed by different groups of people often turn into major obstacles for two-way communication in an efficient manner. Having common understanding also means that terminologies applied by all parties are equivalent. Refinements, modifications and discussions can always be made towards the same terminologies to cope with specific requirements. Hence, the study of encoding the family relations using ontology language is relatively important and meaningful as ontologies are capable of storing family biological relationships more efficiently. In the meantime, ontologies provide shared genealogical vocabularies and common standards for communicating the general genealogical knowledge which address fundamental issues in communicating the knowledge for the same domain among different parties.

This project is beneficial to all humans since everyone has his or her own unique family history. The advantages of this research can be enlarged to support medical fields when proposed Family Ontology is annotated with medical conditions. Therefore, this project is also significant to the healthcare environment since it shows that ontology is capable of building a more powerful and interoperable information system in the medical area. Family Ontology not only helps to store and communicate general family history knowledge conceptually and efficiently, it also supports other domain experts in transferring, processing, reusing and sharing ontology knowledge with other group of researchers. Based on the common standards and terminologies applied within the proposed Family Ontology, discussion among doctors, families and domain experts can be conducted more easily without communication barriers. Wise decisions and conclusions can always be drawn after effective communication and discussion among the key parties.

Since there are previous works available for reuse, an effort will be put on the enhancement of existing works instead of building the proposed ontology from scratch. As such, the main aim of this research will be producing a consistent Family Ontology with other additional features such as reusability, maintainability and inferencing capabilities. Consistent and high quality Family Ontology is always preferable and desirable since it allows effective sharing, transferring and reusing of common genealogical terms to be conducted more easily by all interested parties.

1.2 Problem Statement

Other than storing the family biological relationships, Family Ontology can also be used to support other important functions in different areas. For instance, Family Ontology can be mapped with Medical Ontology to produce Family Medical History Ontology. Family medical histories are very useful for a doctor in accessing the risk of a disease being passed on to their offspring and suggestions of treatments for a particular disease. However, an important prerequisite prior to a robust Family Medical History Ontology is having a precise, consistent, well-designed and complete Family Ontology. Only with a well-structured, consistent and complete Family Ontology, a computer can process, analyze, interpret and acquire the new inferred family knowledge intelligently in a shorter duration. This will definably speed up the diagnosis of a patient and improve the quality of the healthcare systems when a high quality Family Ontology is integrated with Medical Ontology to produce a more complex system.

There are existing genealogical ontologies proposed by different groups of researchers when the semantic web emerged as third generation of the web. However, existing ontologies still lack certain important concepts and properties for the domain of family relations. This may cause ontology to be unable to deliver the full potential of exchanging family history information among family members, doctors and other interested parties. Moreover, the existing ontologies still lack axioms and SWRL rules for consistency checking purposes. Consistency of ontology is fairly important as inconsistent ontology leads to misinterpretation of actual semantic meaning of the data.

Therefore, the objective of this research is to build a new Family Ontology where all required axioms, rules, new terms and properties will be embedded within the resulting ontology to support the requirements of the proposed ontology.

1.3 Objectives

The objectives of this project are as follows:

1. To build a Family Ontology that meets the consistency checking criteria.
2. To evaluate the consistency of the Family Ontology using Pellet, HermiT and FACT++.
3. To compare and analyze the results of consistency checking for the above tools mentioned in (2).

Table 1.1: List of objectives, methodologies and validation methods

Objectives	Methodologies	Validations
1. To build a Family Ontology that meets the consistency checking criteria.	Creating the family reference ontology using the guidelines provided in [11]. The ontology will be developed using the latest Protégé ontology editor version 4.3[12].	1. The validation of the results will be done using the FACT++, Pellet and HermiT.
	Refinement of Family Ontology to confirm to the consistency metric.	
2. To evaluate the consistency of the Family Ontology using Pellet, HermiT and FACT++.	Verification and Validation (V&V) will be done using a framework for ontology evaluation [13]. The V&V will cover the ontology terms, inference rules and instances.	2. The validation will include the resulting new inferred instances through the use of inference rules associated with ontology.
3. To compare and analyse the results of consistency checking for the above tools mentioned in (2).	Verification and validation results using heterogeneous tools are compared and analysed.	

1.4 Scope

For this study, the project will develop a case study involving seventy-one (71) family members for up to three generations of relatives. However, "in-law" relations will not be included in this research. Verification and validation of proposed ontology will focus on the consistency metric only.

1.5 Dissertation Outline

This chapter presents the overview of this research and the impacts of proposed ontology towards other fields. In this chapter, we discuss the problems faced by current approach and how ontology offers a better alternative solution than traditional method. Besides, we listed out some existing genealogical ontologies with similar domain as our reference. We also state the advantages of using ontology language to model domain of family relations and the importance of having a consistent ontology. In spite of these, we also sketched out the objectives, methodologies and research scopes for this research too.

1.5.1 Chapter 2: Literature Review

This chapter provides a comprehensive review on prior researchers' works. This involves extensive comparisons on the existing tools or ontologies which offer the same functionalities as the proposed Family Ontology to be developed. The comparison will focus on the limitations, characteristics, capabilities and features of existing genealogical ontologies. Besides reviewing the internal structure, taxonomy, consistency and completeness of concepts, properties and relations for three existing genealogical ontologies, we also review a list of existing ontology reasoners in terms of their attributes. One out from three existing family ontologies which is closest to the system requirements will serve as the base for customizations.

1.5.2 Chapter 3: Research Methodology

This chapter depicts the methodology being applied in this project which consisted of four sequential steps. The four main phases are strategy design and data, followed by ontology building using relevant concepts, properties, rules and axioms before verifying the consistency of the resultant ontology and the latter ontology refinement if any bugs are discovered in the consistency checking phase. Consistency verification is a fundamental part in ontology development lifecycle since a consistent ontology eliminates false definitions and statements within the proposed ontology.

1.5.3 Chapter 4: Experimental Results and Discussions

This chapter presents the experimental results yielded once the ontology development phases were completed. The discussions revolve around the results of consistency checking using heterogeneous reasoners such as FACT++, HermiT and Pellet. The outputs of consistency checking for different reasoners were captured, compared and analyzed to support the outcomes of this research. This chapter primarily demonstrates how consistency of proposed Family Ontology can be evaluated via different ontology reasoners and how these evaluation results might vary from one another.

1.5.4 Chapter 5: Conclusion

This chapter summarizes all of the research activities that have been done throughout the entire ontology development lifecycle. The contributions of this research are listed and discussed in this chapter. Some of the possible future works are identified in order to enhance and enlarge the scope of this project to support other fields. This allows the improvements of current ontology to be carried out in the coming future in order to cope with the specific requirements of other areas.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

Researchers from all over the world focus extensively on genealogical studies as family history has a very close relationship with human health. According to the definition from Oxford, genealogy is the study of family history including the study of who the ancestors of a particular person were. The efficiency of a healthcare system needs to be improved so that more and more patients can be cured in a shorter duration, without compromising the quality of services at the same time [14].However, effective recommendations on treatments or precautions to patients can only be made when precise, consistent and accurate parental history data is given to a doctor [15, 16, 17]. Hereditary diseases have long attracted public concern. This is because people nowadays have become more health conscious. With family history data, preventions can be taken earlier to minimize the risk of genetic diseases. In order to obtain the family history, there exists a need for a tool that can aid people in constructing their own family tree before medical conditions can be annotated to those family history data.

Previously, there were some good efforts initiated from other researchers in building the applications which assist people in building their own family tree. This includes "*My Heritage Family Tree Builder 7.0*"[18] and *"Family Echo"* [19]. As time passes, researchers realized that there is a communications gap between machines and humans as most machines have been designed to be machine-readable instead of machine-understandable. One of the greatest challenges faced by today's web is a lack

of common standards and shared knowledge among humans and computers in which ontology can be a solution for this fundamental issue. Ontology does not only add semantic meaning to the data contents but also makes the data within the resulting ontology well-connected with each other. In the meantime, it also brings other benefits such as reusability and resource-sharing over distributed systems across heterogeneous applications around the world. Reusability has become the key factor which contributes to a more robust and interoperable system nowadays. This is because a well-structured ontology can always be extended and enhanced easily with slight modifications only towards reusable units by other domain experts. This avoids long hours spent to create a new ontology from scratch.

Knowledge encoded via ontology language for one particular domain can always be published, shared and accessed from other applications through the network which facilitates information exchange. In this way, efforts, resources and time required to build a brand new ontology for one domain can be saved. This is the main reason why semantic web has encountered a quick evolution in recent years for better knowledge representation for a variety of fields as it promotes semantic reasoning, resource sharing and reusing [5, 6]. There are continuous and progressive studies made by researchers on genealogical field nowadays. At the beginning, the main focus was on converting the family tree into ontology format without much consideration of the consistency, inferencing, axioms, rules and constraints. Since there are some existing family ontologies available on the web, restructuring the existing ontologies can be made by augmenting the reasoning capabilities, modular and taxonomy structure. Once the refinement of ontology components for Family Ontology has successfully been completed, consistency checking can be started to ensure that the resulting ontology is consistent and reliable where no contradictory statements can be found within the ontology. In this way, the performance and quality of the developed ontology can be increased. Moreover, ontology with maximal inference can provide intelligent and automated supports which allow the machine to understand the content of Web and generate the data contents automatically instead of explicitly defining every single statement needed.

2.2 Semantic Web

In accordance with Berners Lee [20], semantic web is an extension of the current web in which the existing information on the web are encoded meaningfully and given a well-defined structure, thus enabling computers and humans to communicate in an efficient manner. This indicates that the existing web is facing a transformation from being machine-readable into machine-understandable. In semantic web, all the information has explicit meaning which enabling the machines to interpret, process, infer and derive the new knowledge to support particular mission in real time applications. Therefore, the ultimate goal of semantic web is to create a web of meaning instead of being just a source of reference for a variety of information on the web. Semantic web is capable of providing a common framework that allows the data within the existing web to be shared and reused by heterogeneous applications.

Ontology is the backbone of the semantic web that contributes to the knowledge sharing over the distributed systems. Interoperability issues for different applications in different organizations can be solved when a shared framework for particular domain is created. Several existing ontologies built previously can be shared and accessed according to the needs of domain experts. Recently, semantic web has been accepted widely for better knowledge representation, communication, sharing and reasoning on the web. As semantic web is growing rapidly now, numerous languages that support the functionalities of ontology have been invented. Web Ontology Language (OWL), Resource Description Framework (RDF) and Resource Description Framework Schema (RDFS) are among the basic representation languages for Semantic Web. Examples of applications that make use of ontology include e-commerce websites or enterprise websites and search engines (http://swoogle.umbc.edu/) and multi-agent. The purpose of multi-agents is to provide a shared understanding of domain knowledge and allowing for easy communication between agents. Figure 2.1 shows the semantic layer cake proposed by Tim Berners-Lee which consists of rule layer, ontology vocabulary, logic, RDF and RDFS schemas.

Semantic Web is also known as third generation of the web. Semantic Web has made the step towards "Knowledge Web" that concentrates on machine processable

meaning of information. The "Knowledge Web" enables the machine to interpret, understand the information and process it in an intelligent way instead of connecting other pages via predefined hyperlinks in HTML. This has led to the evolution of Semantic Web. With "Knowledge Web", intelligent services such as search agents, multi-agents, information brokers and information filters are facilitated. Knowledge serves as the basis for data manipulation and reasoning. Knowledge representation is the field of artificial intelligence that represents the knowledge symbolically to facilitate manipulation of knowledge by reasoning method in an automated way [21]. Nowadays, semantic web has gained popularity as it overcomes the communication gap between humans and machines by adding the semantic or meaning to the data in machine understandable and process able format.

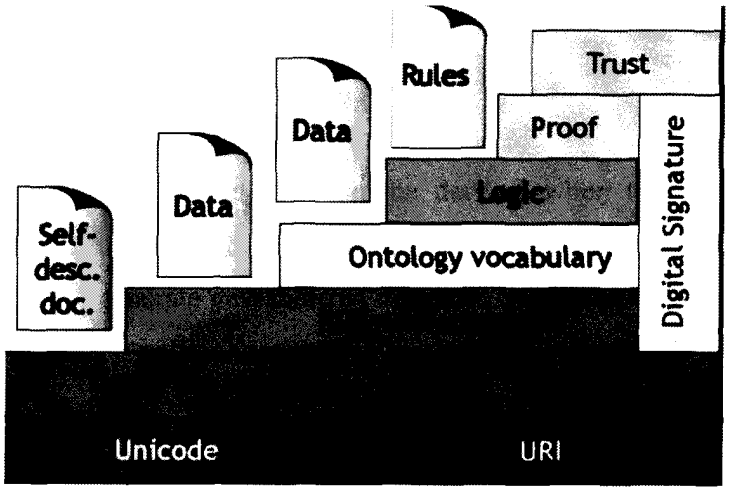

Figure 2.1: Semantic Web layer "cake" by Tim Berners-Lee [22]

2.2.1 Reasoning on Semantic Web

Semantic Web is an extension or evolution of the current web in which it can provide an efficient reasoning support over the data represented in ontology language. Evolution of Semantic Web is driven by two main goals: (i.) to interpret and understand the semantic meaning of huge data residing on the web and (ii.) to infer new knowledge automatically

from existing facts. According to Long [18], progression of medical informatics is strongly influenced by the development of various reasoning methods in the world. These reasoning methods work by organizing various relations occurring within the medical domain. The relations may include associations, probabilities, causality, functional relationships, temporal relations, locality, similarity and clinical practice. In the past, decision support systems are widely used in the healthcare environment to assist doctors in diagnosis based on some pre-conditions. Decision tree, rule-based reasoning (RBR), Bayesian probabilistic and case based reasoning (CBR) are the backbones used to support the expert systems. However, inability of the machines to understand the meaning of the data often leads to misinterpretation of the actual semantic meaning of the data contents. Machine learning-based systems have emerged specially to deal with semantic meanings of data on the web. Semantic Web has appeared as an alternative solution for the expert system whereas ontology serves as the backbone.

Ontology is capable of providing inferencing and reasoning capabilities where additional facts from the present data can be derived when the rule languages are implied on a reasoner. Reasoner or inference engine is a fundamental component of ontology as it helps to generate new inferred instances from asserted axioms intelligently against mass of data in an application. By applying equivalence, transitive, inverse, subclass and disjoint, other new knowledge such as similarity concepts, superclasses and subclasses relationships in both directions can be discovered indirectly. However, automated reasoning cannot be performed against inconsistent ontology. Any inconsistent modules must be sorted and removed before reasoning can be invoked successfully. This requires an ontology reasoner to deal with the reasoning tasks. New entities can be classified based on its types, super and sub type relationships or equivalent when consistent ontology is provided.

Reasoning is normally performed via Description Logic (DL) where it is often used to reason about objects or classes. Some of the famous reasoning engines are FACT++, Pellet, HermiTandKAON. These reasoners are used to infer the implicit meanings of classes, properties and individuals. All of the reasoners stated above are currently under active development and improvement to ensure the new features are

ready for users. OWL realizes the inference goal by introducing class subsumption, property subsumption, inverse properties, equivalent properties, symmetric properties and transitivity properties. Inferencing is useful for enormous data since defining every single fact explicitly in an ontology can be very troublesome and time consuming. Errors are often caused by human mistakes when there are larger number of facts needed to be declared manually instead of being machine-generated. Machine-generated facts using predefined rules are correct provided that the rules inserted are true. Hence, the rules applied within the proposed Family Ontology should be examined carefully to make sure the rules are correctly declared in the right position.

2.2.2 Description Logic

Description logics are subsets of the first order logic and primarily used for knowledge representation. Knowledge representation based application is normally composed of two main components namely TBox and ABox. TBox denotes the terminology of application domain. Terminology is composed of concepts and roles. ABox contains assertions of individuals. For instance, an individual is declared to be instance of a concept which resides in ABox. Figure 2.2 shows the architecture of a description logics knowledge-based system.

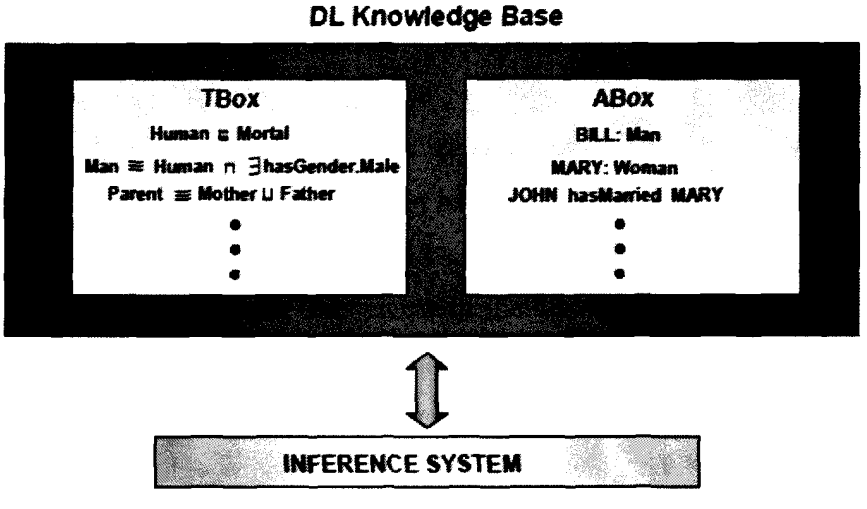

Figure 2.2: Architecture of a Description Logics knowledge-based system [19]

2.2.3 Rules

Desired expressivity features or reasoning can often be observed after embedding appropriate rules into the applications or machines. Rules are invented to enable new knowledge to be inferred intelligently by machines itself instead of explicitly defining every single statement required for a particular domain. New knowledge derived via intelligent agents solves fundamental issues such as human mistakes, resource and time wasting when a large number of facts are to be stated explicitly. Human mistake usually happens when they are forced to deal abundant data manually. Intervention of machines is always preferable and recommended in computer science as the probability of human mistakes can be minimized tremendously. Rule languages applied within the resulting ontology will be processed by a reasoner. A reasoner is usually invoked to accomplish the inferencing task. The most popular rule language applied by most of the domain experts is SWRL. Rules plays a major role in Semantic Web since they provide intelligent and automated support and this fact has been agreed upon by Tim Berners-Lee [20].

2.2.3.1 SWRL Rules

SWRL is a rule language recommended by W3C in 2004 which is used to express rules using RDF for Web context. Specifically, SWRL is the combination result of Web Ontology Language (OWL) and also Rule Markup Language (RuleML). SWRL rule supports the reasoning and inferencing capabilities of ontologies by introducing conditional rules or logical rules which can be applied within the ontologies. Two main parts for SWRL rule are antecedents and consequents. If the statement declared in antecedent part is true, then the statements in the consequent part will be applied. In "*Aunt Rule*", if a person's father has a sister then that sister can be treated as the person's aunt. There are some advantages offered by SWRL. This includes the simplicity it offers and SWRL is compatible with OWL. Therefore, applying SWRL rules in ontology is

much easier. Moreover, most of the inference engines such as Pellet and HermiT have been designed to support this rule language specifically.

2.3 Ontology

Ontology is the key component of semantic web. It is used primarily to describe the contents of web or capture knowledge about some domain of interest by semantic means. An ontology makes use of classes, relations and instances as definitions to describe the concepts for one phenomenon and properties to model the relationships which hold between those concepts via formal relations. Ontology provides the common vocabularies for publishing data and allows discovery of its contents via other applications over distributed network. In accordance to Gruber [5], ontology is a formal, explicit specification of a shared conceptualization about a particular domain. Gruninger and Fox [25], on the other hand, define ontology as a formal description of entities and their properties, relations, constraints and behaviors. One of the most attractive features of ontology is its capability that facilitates knowledge processing, sharing, reusing and information exchange between different parties through heterogeneous web applications services as ontology is able to provide a common understanding about a particular domain. Ontology is able to supply a richer description using terms and relationship between them in the application domain. Developing ontology from scratch is still a common trend for most developers nowadays but it has led to many of the potential and relevant knowledge not being reused wisely.

Ontology reuse can be defined as the process of reusing parts of or whole ontologies in order to support specific application requirements. Ontology reuse is preferable since it is a key factor that contributes to high quality and cost-effective ontologies. Ontology reuse is recommended as it reduces the cost, time and resources incurred while developing a brand new ontology. Moreover, huge amounts of ontologies with same domain from different research projects are available publicly. It is always recommended to modify the existing ontology directly whenever there are high similarities where most related concepts exist in previous ontology [26].There are few

criterias to be considered when choosing appropriate existing ontology. These criteria include high similarity, correctness, reusability, interoperability and maintainability. Correctness is related with accuracy, consistency and completeness of an ontology. An ontology is easier to be maintained when the ontology possesses the characteristics such as simple, concise and modular architecture. Ontology can be extended easily when the ontology is of high generality, simplicity, modularity and independence [27].Hence, existing ontology should be chosen based on all the attributes above to ensure the ontology built is of high quality.

Building an ontology of high quality contributes to a higher degree of reuse, lower maintenance cost and better cooperation between humans and computers. Well-designed ontology should be intuitive to human users, expressive enough and completed with intelligent reasoning support. This includes clear syntax which enhances the readability by human users and formal semantics used are understood by machines to facilitate interpretation and analysis by intelligent agents.

2.3.1 Ontology Languages

Various ontology languages have been designed for supporting ontology modeling. This includes RDF, RDFS and OWL. Figure 2.2 above shows the latest Semantic Web Stack Diagram. From this diagram, we can clearly see that RDFS is located on top of RDF. This also means RDFS is extending RDF whereas OWL is extending RDFS and RDF. OWL extends both ontology languages by adding more features such as reasoning and richer set of vocabularies are supported.

(A) RDF

RDF is a Resource Description Framework. Its primary purpose is to describe the resources on the web using named properties and property values. It is a graph-based data model with labeled nodes and directed, labeled edges. The building blocks of RDF

are RDF statements, which corresponds to the edges in the graph. A RDF statement is composed of three components namely subject, predicate and object as illustrated in Figure 2.3 below. The subject represents the source of the edge and it is used to identify things or concepts whereas the object of a statement is the target of the edge or value of property. The predicate of a statement on the other hand denotes the kind of relationship in between a subject and an object. Predicate denotes property or characteristics of the subject. Normally, assertions are based on the triple form <subject, predicate, object>.

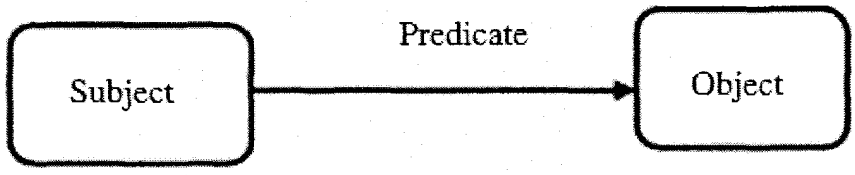

Figure 2.3: RDF graph structure

(B) RDFS

RDFS is a Resource Description Framework Schema that provides rich vocabularies to be used in RDF graph. RDFS is needed to add meaning to the data as RDF is just a data model that doesn't convey any significant semantics. Some of the RDFS schema which are usually applied to define classes and properties are rdfs:Class, rdfs:subClassOf, rdfs:subPropertyOf, rdfs:domain and rdfs:range. RDFS allows representation of classes and properties in hierachical structure and also offers domain and range restrictions on the properties. RDFS is useful since it is capable of providing inference based on declared schema. New derived knowledge will be generated once inference is run on the data provided. For instance, rdfs:subClassOf and rdfs:subPropertyOf are responsible in generating taxonomies and hierachies among properties and resources.

(C) OWL

OWL is a standard representation language proposed by W3C to encode the knowledge of Semantic Web in February 2004. In fact, OWL is the extension of RDF and RDFS

where both of them are early Semantic Web standards endorsed by W3C. OWL consists of entities and axioms in general. It is used to specify classes and properties in a form of description logic. There are three types of entities which exist in OWL ontology which are classes, properties and individuals. Individuals refer to objects of the domain knowledge; properties are used to link those individuals via binary relations whereas classes are set of individuals with common characteristics. In this case, family members represent the individuals whereas family biological relationships represent the properties in the family domain and Person is the class for all instances defined in Family ontology. OWL provides a richer set of class operators analogous to Boolean operators which are not supported by previous standards such as intersection, union and negation. OWL offers greater machine interoperability than XML as additional vocabularies and formal semantics are supplemented. OWL extracts the strengths of Description Logics, besides using well defined semantics; it also supports reasoning [19].

2.3.2 Ontology Evaluation Criteria

There are two main ontology evaluation criteria to be checked which are stated in [13]. Two ontology evaluation criteria to be fulfilled are ontology completeness and consistency. Both criteria are used to reveal human mistakes such as inaccurate or incomplete definitions of ontology components, incorrect classification of instances or concepts to the wrong types, super type or sub type relationships. Ontology which is incomplete, inconsistent and inaccurate will indirectly cause the resulting ontology to be unable to perform the functionalities as expected since the machine could misinterpret the actual meaning of the data content.

2.3.2.1 Consistency

According to Gomez Perez [13], consistency checking should exclude any contradicted definitions for all ontology components that have been defined within an ontology.

Important ontological components include concepts, properties, rules, axioms and instances. This also means that contradicted rules, classes, properties, instances or axioms are strictly prohibited within the proposed Family Ontology. Consistency checking can be aided by heterogeneous tools available on the web. There are a variety of existing tools which support consistency checking which are HermiT, Pellet and FACT++ [28]. However, only three tools are applied in this research and the results of evaluation are captured, analyzed and compared in Chapter 4.

Consistency checking covers many areas of checking which include membership checking, instance checking and relationship checking. All these areas must be checked carefully in order to ensure conflicting elements do not exist in the proposed Family Ontology. Precise and accurate ontology is crucial in facilitating efficient knowledge sharing among key parties across distributed systems. Inconsistent ontology often causes unwanted errors or conflicts to happen within an ontology which leads to misinterpretations of the actual semantic meaning of a particular ontology by underling machines. Misunderstanding of actual semantic meaning of data contents may result in wrong feedback being reverted to users. The ontology reasoner is invented to overcome the consistency issues within the resulting ontology specifically besides deriving new facts from existing knowledge. One of the important functionalities of the reasoner is to check the consistency of ontology besides performing reasoning task. Some of the famous reasoner tools are HermiT, FACT++ and Pellet. To check the consistency of an ontology, a reasoner must be selected first before proceeding to the reasoning task. The details of consistency checking is discussed in Chapter 4.

2.3.2.2 Completeness

According to Gomez, an ontology is considered as complete when all expected knowledge that should be present in the ontology can be found either as asserted statements or inferred statements within the ontology [13]. Grüninger and Fox suggested that ontology completeness can be achieved via proving the theorems through a list of competency questions [25]. A set of queries can be taken as questions in which the

proposed Family Ontology must be able to answer. These queries can act as the requirements for the system and validate the completeness of the resulting ontology. In fact, axioms and rules can also be declared within the resulting ontology to ease the completeness checking process. Declaring a rule and executing the rule via inference engine will derive other implicit facts. Those facts are accurate provided the rules declared are also correct.

Since the scopes of this project are three generations of family relations, the proposed Family Ontology is considered as complete when all of the family relations under three generations of relatives are included within the proposed ontology. These include:

i.) Concepts which fall under three generations of family relations. The concepts can either be explicit concepts or implicit concepts.

ii.) Properties which fall under three generations of family relations. The relations can either be explicit relationships or implicit relationships.

iii.) Instances which belong to three generations of family relations. The instances can either be explicit instances or implicit instances.

2.3.3 Ontology Reasoners

Ontology reasoners or inference engines are mostly used to derive new facts from pre-existing knowledge. Other than deriving new inferred statements from known facts, reasoners can be utilized to check the logical consistency of the ontology model. There are various types of reasoner tools proposed by many researchers in the last few years. Some of the popular reasoners are Pellet, HermiT, RACER, FACT++ and ELK. Each of these tools contains unique attributes and characteristics which make them vary from one another. Most of the reasoners apply first-order predicate logic to perform the reasoning tasks.

(I) HermiT

HermiT is an open source reasoner tool designed specifically for ontologies inference and it is written using Web Ontology Language (OWL). Similar to most of the other reasoners tools available publicly or commercially, the basic functionalities of HermiT are to examine the consistency of resulting ontology and identify the subsumption relationships within classes or properties. HermiT is the first OWL reasoner based on hypertableau calculus which is capable of providing a speedy and high performance in reasoning compared to previously used algorithms. For now, the latest version of HermiT 1.3.8 has been released under the GNU Lesser General Public License (LGPL). HermiT is able to classify a number of ontologies which had previously proven too complex for any available system to handle within few seconds.

(II) FACT++

Fast Classification of Terminologies or FACT++ is another new version of the well-known FACT OWL DL reasoner developed in the past. The latest version of FACT++ is 1.6.3 which was released recently in 30 May 2014. There were some improvements over the previous release. FACT++ utilizes the preexisting algorithm as in FACT but with minor changes in internal architectures. FACT ++ is another DL reasoner implemented in C++ language in order to generate an efficient software tool and maximize portability. However, FACT++ supports OWL DL and OWL 2 DL partially. This tool is based on optimized tableau algorithms for general TBoxes and incomplete support for ABoxes. It is an open source software for SHOIQ(D). One of the disadvantages of FACT ++ is probably inefficiency in supporting complete ABox reasoning. Therefore, FACT ++ is often discouraged for applications which require the functionalities such as instance classification and retrieval.

(III) Pellets

Pellet is another open source OWL-DL reasoner implemented in java by The Mind Swap Group. Pellet is designed specifically to handle expressive OWL ontologies. Similar to FACT++, Pellet is based on optimized tableau algorithm and able to support expressive description logics. Pellet is the first complete OWL-DL reasoner tool that provides good support for all OWL DL SHOIN (D) and debugging facility which facilitates error discovery for inconsistent ontologies. Compared to other reasoners which are also able to detect the inconsistencies between concepts of the domain, Pellet can provide explanations and justifications of reason a concept leads to dissatisfaction. This way, users will understand the actual problem which causes any inconsistencies. This reasoner is augmented with some additional features. Extra features which are supported by Pellet include Unique Name Assumption – UNA, closed world reasoning and SPARQL query. Figure 2.4 shows the main components of the Pellet reasoner.

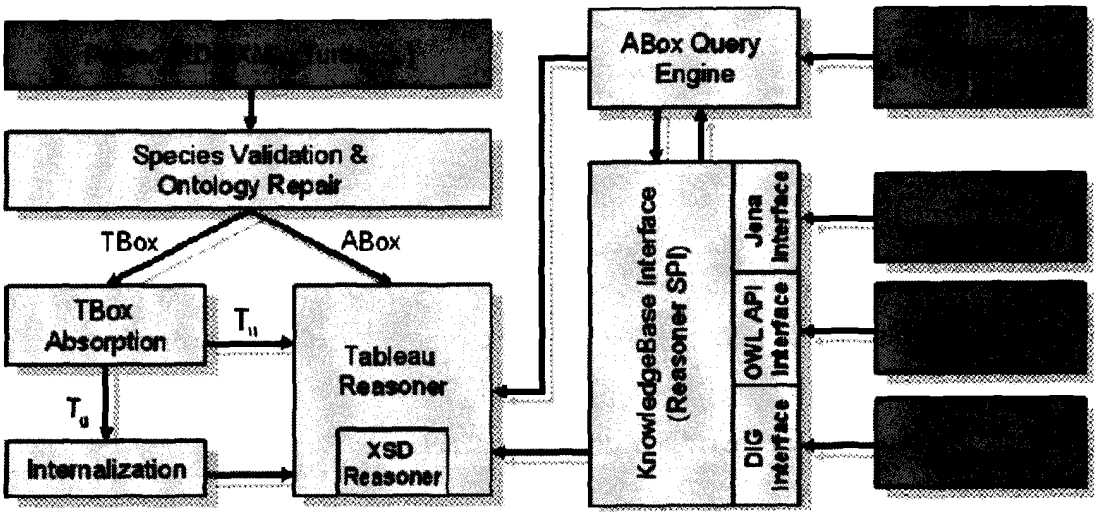

Figure 2.4: Main components of the Pellet reasoner [29]

Three reasoners which have been chosen and reviewed in this research are Pellet, FACT++ and HermiT. These reasoners are analyzed, reviewed and compared from different perspectives. Ontology properties reviewed in the following sections are

methodology, soundness, completeness, expressivity, native profile, incremental classification, rule support, ABox reasoning and other features. Pellet and FACT++ employs the same reasoning algorithm that is Tableau for general TBOX (subsumption, classification and satisfiability) and ABOX (retrieval, conjunctive query answering) whereas HermiT applies Hypertableau as its underlying algorithm for inferencing purpose. Soundness and completeness evaluate whether all possible inferences are inferred or not. All tools being reviewed below provide complete possible inferences. All reasoners can support SHOIQ (D). SHOIQ as an extension of description logics (DLs). SHOIQ is can provide a more expressive means and it is designed to compromise between expressivity and computational or complexity in reasoning. FACT++ and HermiT employ DL as their native profile whereas Pellet is based on DL and EL. The primary purpose of DL is to add expressivity in language whereas EL aims to provide scalable reasoning in TBOX.

Rule language was invented and incorpoted into the reasoner tool to increase expressivity of ontology. SWRL is the most popular rule language due to the simplicity it offers compared with other rule languages. SWRL in DL-Safe Rules notion is supported by Pellet and HermiT but not FACT++ which also means that those rules will only be applied to named individuals in the resultant ontology. Therefore, new inferred instances cannot be derived when these rules are reasoned via FACT++. An attractive feature from Pellet is it can be used to support incremental classification which is not allowed in FACT++ or HermiT. Incremental classification allows Pellet to compute the inferred hierachy for affected modules only when addition or removal operations have been done by users. This significantly increases the performance of Pellet.

FACT++, HermiT and Pellet are not restricted by the users's operating systems.These tools can function well regardless of the operating systems of users employ. Among all the reasoners presented in this study, only Pellet supports justications for any inconsistency and conflicting error. Three ontology reasoners reviewed in this study support ABOX reasoning that is reasoning with individuals such as instance checking. Among the three reasoners, only Pellet can work well with Jena API. Pellet is an open source reasoner implemented in Java. FACT++ is another open source C++ based OWL-DL reasoner. Similar to both ontology reasoners stated above,

HermiT is also an open source Java based reasoner which can be manipulated and accessed by anyone who wish to perform the reasoning tasks. All the reasoners above are categorized under open source tools supported by Protégé ontology editor. The user just need to choose a reasoner and invoke it subsequently. The outputs will be the inference results. Table 2.1 below summarizes the comparison of three ontology reasoners mentioned above.

Table 2.1: Comparison of three ontology reasoners [30]

Tools / Review features	Pellet	FACT++	HermiT
Methodology	Tableau based	Tableau based	Hypertableau based
Soundness	Yes	Yes	Yes
Completeness	Yes	Yes	Yes
Expressivity	SROIQ (D)	SROIQ (D)	SROIQ (D)
Native Profile	DL, EL	DL	DL
Incremental Classification (Addition, Removal)	Yes	No	No
Rule Support	Yes (SWRL)	No	Yes (SWRL)
Platforms	all	all	all
Justifications	Yes	No	No
ABOX Reasoning	Yes	Yes	Yes
Jena Support	Yes	No	No
Implementation Language	Java	C++	Java
Availability	Open source	Open source	Open source
Protégé Support	Yes	Yes	Yes

2.3.3.1 Importance of Reasoners

The major responsibilities of reasoner include checking the logical consistency of the ontology components, maintaining a class hierarchy, classifying an entity within an ontology and making queries towards the resulting ontology. According to W3C, reasoners are usually used to create the taxonomy structure within ontology and support

REFERENCES

1. Reid, G. and Emery, J. *Chronic disease prevention in general practice - applying the family history.* 2006. vol. 35, no. 11, pp. 879 – 885.
2. Yoon, P.W., Maren, T. S., Kris, L.P., Marta, G., Andrew, F. and Muin, J. K. *Can family history be used as a tool for public health and preventive medicine?* 2002. vol. 4, no. 4, pp. 304 – 310.
3. King, R. A., Rotter, J. L. and Morulsky, A. G. *The Genetic Basis of Common Diseases*, Oxford University Press, New York, NY, USA. 1992.
4. Baltimore, M. D., Williams and Wilkins. *Guide to Clinical Preventive Services.* 2rd ed. U.S. Preventive Services Task Force. 1996.
5. Gruber, T. *A Translation Approach to Portable Ontology Specifications.* 1993. vol. 5 no. 2, pp. 199-220.
6. Guarino, N. *Formal Ontology and Information Systems.* Proceedings of FOIS'98, Trento, Italy. Amsterdam, IOS Press, pp. 3-15. 1998.
7. Smitha, B. and Ceustersb, W. *Ontology as the Core Discipline of Biomedical Informatics.* Forthcoming in Computing, Philosophy, and Cognitive Science, G. D. Crnkovic and S. Stuart (eds.), Cambridge: Cambridge Scholars Press. 2006.
8. Bodenreider, O., Smith, B., Kumar, A. and Tolksdorf, R. A. Investigating subsumption in SNOMED CT: an exploration into large description logic-based biomedical terminologies. Artif Intell Med. 2007; 39: 183–195.
9. Smith, B. and Kumar, A. *On the Application of Formal Principles to Life Science Data: A Case Study in the Gene Ontology. In: Proceedings of DILS 2004 (Data Integration in the Life Sciences), (Lecture Notes in Bioinformatics 2994), Berlin: Springer; 79-94.* 2004.

10. Golbeck, J., Fragoso, G., Hartel, F., Hendler, J., Oberthaler, J., and Parsia, B. *The National Cancer Institute's Thesaurus and Ontology*. Web Semantics: Science, Services and Agents on the World Wide Web. 2003. 1(1):75–80.
11. Muthukkaruppan, A. and Leon, S. *Guidelines for Constructing Reusable Domain Ontologies*. International Conference in Autonomous Agents and Multi-agent Systems Workshop on Ontologies in Agent Systems, Melbourne. 2003.
12. Protégé 4.3. (2000). *Ontology editor*. Retrieved May 16, 2014 from http://protege.stanford.edu.
13. Gomez, P. A. *Ontology evaluation*. In Steen Staab and Rudi Studer, editors, Handbook on Ontologies. 2004. First Edition, chapter 13, pages 251-274. Springer.
14. Mikkola. *Hospital Pricing reform in the public health care system- an empirical case study from Finland*. 2003. International journal of Health Care Finance and Economics, 3 (4) 267-286.
15. Wood, M. E., Stockdale, A. and Flynn, B. S. *Interviews with primary care physicians regarding taking and interpreting the cancer family history*. 2008. *Family Practice*, vol. 25, no. 5, pp. 334-340.
16. Fuller, M., Myers, M., Webb, T., Tabangin, M. and Prows, C. *Primary care providers' responses to patient-generated family history*. 2010. Journal of Genetic Counselling, vol. 19, no. 1, pp. 84–96.
17. Suther, S. and Goodson, P. *Barriers to the provision of genetic services by primary care physicians: a systematic review of the literature*. Genetics in Medicine, vol. 5, pp. 63–65. 2003.
18. My Heritage Family Tree Builder 7.0. *Family Tree Builder*. Retrieved October, 2014 from http://www.myheritage.com/.
19. Family Echo. *Family Tree Builder*. Retrieved October, 2014 from http://www.familyecho.com/.
20. Tim, B. L., James, H. and Ora, L. *The semantic web*. Scientific American. May issue. 2001.
21. Ronald, J. B. and Hector, J. L. *Knowledge Representation and Reasoning*. Elsevier, ISBN 978-1-55860-932-7, pp. I-XXIX, 1-381. 2004.