Chapter 7: Conclusions and Future Work

This chapter summarizes the work presented in the thesis and the conclusions reached. The main goal of the thesis is to provide an approach that integrates a knowledge base system with an information system while retaining the flexibility for each of the component system to function on its own. A framework has been provided that allows the modeling of knowledge base system using a new UML Profile, addition of new formalisms using the process of knowledge map for the implementation of the ontologies, designing and implementation of new neurological databases and finally the integration of ontologies into databases. This approach is useful in many applications such as health informatics and neuro-informatics.

In the following sections, the summary of the work is presented in section 7.1 while the principal contributions of the research are discussed in section 7.2.. Finally, the recommendations for future work are discussed in section 7.3.

7.1 Summary

The contents and the achievements of the work has been summarized in four categories:

- creation of new, real-time clinical databases (section 7.2.1)
- ontological integration of heterogeneous data sources (section 7.2.2)
- new UML Profile created on the modeling of DAML+OIL based ontology on People With Parkinson's Disease (PWPD) (section 7.2.3)
- knowledge based systems (KBS) developed for the different facets of PWPD (section 7.2.4).

7.1.1 Creation of new, real-time clinical databases

The absence of any clinical databases on PD precluded the efficient inference of knowledge. As such, it was warranted that a clinical and a physiotherapeutic database be developed that will act as the source to give critical insights in the advancement of medical knowledge for PWPD.

The clinical database (PDCDB) has been designed to store clinical information that will provide medical researchers and practitioners alike with new knowledge into the physiological and socio-economic factors affecting the drug concordance of PWPD. This has been made possible by designing the database into the following 6 categories: (a) Activities of Daily Living, (b) Drug Concordance, (c) Stages of Parkinson's Disease, (d) Parkinson's Disease Symptoms, (e) Side Effects and (f) Services and support. The database has been implemented on a semantically enabled web site that will allow other researchers to co-relate any work conducted by the fraternity of researchers.

The physiotherapy database (PDPHDB) is an annotated database on the physiotherapeutic aspect of the disease. Annotation was done to provide a description of the database that will provide the users with easy to read retrieval of data. Like the PDCDB, it has been classified into 5 categories: (a) Techniques, (b) Services, (c) Assessment and Carer Needs, (d) Communication and (e) Carer and Physiotherapists Information, the essential components in the arsenal of a physiotherapist.

7.1.2 Ontological integration of heterogeneous data sources

Using the clinical and the physiotherapy data model, two heterogeneous data sources, ontologies were extracted from the underlying schema containing entities, attributes and relationships. This was possible since the logical layer of the data model contains the conceptualization of the specific domain. As such, the explication of the concepts and the corresponding attributes were transferred by reverse engineering the schema. Although integration into ontologies was found to be an iterative process, it expedited the fashioning of the knowledge base in the preliminary stages since the semantics for the domain were already established in the conceptual layer of the database schema.

The entities were transferred as concepts and the attributes of the entities were generated into the properties for the concepts. The ontology generated for PDPHDB contained 26 concepts, 171 relations (or slots) and 50 axioms while that for PDCDB produced 27 concepts, 312 relations and 50 axioms. Relational data models contain a high number of redundant relationships, as such the ratio of classes to attributes

generated is 1:7. Since data modelers use concrete data types to represent entities for the specified domain, the concepts were generated as concrete in nature while the data types for the attributes were translated as string or integer.

It was also found that the database schema provided suggestions in the development of new concepts. Some of the attributes of the concepts were, however, redundant since it did not provide any information in the conceptualization of the domain. This was particularly true during the generation of referential key integrity constraints into attributes that caused confusion to the domain experts during the refining stage of the ontology. The attributes generated were of two types, notably DatatypeProperties corresponding to columns while columns containing referential integrity constraints generated ObjectProperties.

The quality of the extracted ontologies was improved for the purpose of reusability by aligning for the first time, medical ontologies with the foundational ontology of DOLCE to provide ontological commitment. This was done by the linking of the DOLCE ontologies by the subsumptional relationship to the ontologies of PDPHDB and PDCDB.

7.1.3 Creation of UML Profile

The maintenance and administration of any Knowledge base system (KBS) is an integral part in the management of any knowledge based enterprise system. The process of constructing a knowledge-based system is similar to the modeling constructs of the conceptual layer of the software system. UML, as the industry standard has been employed in the design and analysis phase of the software industry. It can help knowledge modelers to structure the blue print of the KBS while allowing knowledge administrators to maintain the KBS. This is possible by using the profile mechanism.

Development of the profile involved mapping software paradigm to ontological constructs. New modeling constructs on neurological knowledge bases have been developed. The modeling of the components of PD knowledge base comprising Symptoms and drugs, dietary, physiotherapy and Speech and language therapy were achieved by the extension of UML profile. The semantics of DAML+OIL were

mapped to the vocabulary of UML. Using the lightweight extension of UML, stereotypes and tagged values were entered to map the modeling primitives of UML to DAML+OIL. The class diagram was employed to model the structure of the conceptual knowledge of each of the knowledge bases. Knowledge modeling has been shown to be useful during the period of knowledge acquisition. This transitory period facilitated the extraction of tacit to explicit knowledge from domain experts. This was possible since the complexities of modeling knowledge were mitigated by visual demonstration, which is not possible with frame based modeling tools like Ontoedit and Protégé 2000. Besides, this also ushered in discussion among domain experts that opened up new vistas of knowledge discourse.

The KBS representation for the Speech and language therapy in UML contained three packages, viz. Communication, TherapyApproches and Swallowing. These were then transferred as DAML+OIL ontology. Each of the packages had an abstract class declared. This was intentional since any new knowledge developed can be added at any time by simply instantiating the abstract class.

7.1.4 Knowledge based systems for PWPD

The process of knowledge map was employed to develop the KBS for Parkinson's Disease Knowledge map (PDK). The KBS of PDK consisted of the knowledge base for Symptoms and drugs, the knowledge base for diet, the knowledge base for physiotherapy and the knowledge base for Speech and language therapy. The development was a step-by-step iterative process used to harness the knowledge resources of domain experts. This process was employed in the knowledge acquisition phase by outlining initially the business process map, followed by extracting the knowledge from both direct and indirect sources that is manifested in the structural creation of ontologies, and subsequently by the addition of profiling knowledge whereby relationships like symmetric, transitive and disjoint between concepts are established. Subsequently the knowledge was verified and validated with the help of domain experts.

The knowledge base was enhanced by the addition of axioms. The creation of rules or axioms was made possible by the use of F-Logic. These rules were also used in other knowledge bases within the same domain. Rules such as hyponyms and hypernyms were reused in other knowledge bases. They were used to maintain the integrity of the knowledge base and permit the knowledge base to function within the constraints of the business logic for the domain.

This process also helped to identify potential gaps in knowledge and reduced the burden on experts by sharing knowledge. It was also found that considerable social capital needed to be expended in order to facilitate the process. As such, a facilitator can act as the bridge between the domain experts and the knowledge engineers in order to expedite the development process. Knowledge merging and mapping was performed after the knowledge validation step but it also needed to be re-validated after it was merged

The mapping of the knowledge was performed manually and merging was done by a semi-automatic process. The manual operation was tedious and time-consuming. Although the merging of the KBS was semi-automatic it gave rise to substantial conflicts which had to be resolved by further expert human input. A federated knowledge base that interacts with other loosely coupled knowledge bases was created out of this merging process.

7.2 Conclusions

In this section the principal contributions of the thesis are presented as follows:

- The thesis is primarily concerned with the problem of integrating a knowledge management system for Parkinson's Disease (PD) with heterogeneous data sources. In particular it addresses the problem of extracting ontology from disparate data sources, since the conceptual layer of the data sources capture the semantics of the domain experts, and then aligning with a foundational ontology.
- A problem arose in the generation and extraction of ontologies from relational databases. The parent child table relationship in the data definition language

(DDL) did not get transferred into a subsumptional relationship between the concepts in the ontology.

- The abstraction of some of the attributes was found to be inconsequential and redundant. This was particularly in the case of referential integrity constraints. Although they maintain the consistency of the database, the attributes generated in ontologies did not provide information about the conceptualization of the domain. They only added clutter and had to be removed manually.
- The quality of the ontologies generated and extracted from both the clinical and the physiotherapeutic databases needed to be improved. This was done by aligning the extracted ontologies with a foundational ontology like DOCLE that has a rich set of axiomatised foundational ontology. The benefits of aligning the ontology lay in reusability and facilitating ontological commitment.
- Ontology was used to generate the physical layer of the database. Concepts were generated as tables while the attributes of the concepts were mapped to the corresponding column names.
- The sub-concepts did not map to the corresponding child table but were generated as separate entities. Although the physical layer of the database needs to be improved, it provided the basis for semantic integration of the various components of the system.
- In order to facilitate the co-ordination between the research fraternities associated with PD, two databases were created for the user community to access the data and draw inferences accordingly.
- These databases serve as the repository for the clinical and the physiotherapeutic aspects of PD. The conceptual layer of the schemas was used to extract ontologies as mentioned earlier. The databases serve as the back-end for the semantically enabled web site.
- This research also formulated a novel method to model the DAML+OIL enabled knowledge management system for PD. Since UML is an industry standard and the similarities between object-oriented programming and ontology are many, as discussed in chapter 2, it is being used since the model developed would serve to integrate both the information and the knowledge system for PD.
- This was made possible by using the extensions of UML Profile. The primitives of DAML+OIL for the different knowledge bases like physiotherapy and diet

were mapped to the vocabulary of the UML. The structure of the conceptual knowledge for each of the knowledge bases was modeled using the class diagram.

- The classes in UML were transferred as DAML+OIL concepts while the attributes were transferred as DAML+OIL Datatype Property. The relationships between the classes were mapped to the DAML+OIL Object Property. Concepts belonging to the same class were modeled using the stereotype daml:sameClassAs while disjoint relationships between classes were expressed as daml:disjointWith.
- Theses classes can also be used to generate classes for object-oriented programming languages like Powerbuilder and Java.
- The major benefit of modeling was that it helped both the domain experts and the knowledge engineer in the knowledge acquisition phase due to the simplicity and the graphical representation of both the information and the knowledge system.
- The process of knowledge map was introduced into the formation of knowledge base systems for the symptoms and drugs, physiotherapy, speech and language therapy, and dietary aspect for PD. This process enabled the extraction of both the tacit and the explicit knowledge of PD.
- The knowledge base was enhanced by the creation of F-logic based axioms. The benefit of using axioms is that it enables inferences to be drawn on the knowledge base.
- The knowledge bases were verified and validated in consultation with the domain experts. Any gaps in the knowledge were identified while any redundancies were eliminated. This process also helped to establish a valuable knowledge network among domain experts.
- The use of knowledge merging and mapping as one of the key steps in knowledge mapping was employed. Mapping was done manually while merging was done semi-automatically. While the manual operation was time-consuming, there were substantial conflicts in the semi-automatic process that had to be resolved by further input from domain experts.

The next section in this chapter describes the possible work that can be conducted in the future.

7.3 Future Work

The recommendations made for future work are categorized into four parts: (a) modeling of neurological knowledge base, (b) data modeling of neurological data sources, (c) mobility of data and (d) development of smart clients

7.3.1 Modeling of Neurological Knowledge Base

The modeling developed for PD KBS with UML Profile is new and computationally efficient. The use of UML lightweights like stereotypes expedites the modeling constructs of ontology. However, the creation of axioms necessary to enhance the knowledge base was not possible with the UML Profile on the class diagrams. The following suggestions can possibly accelerate and open up new vistas in the modeling of knowledge bases:

(a) The logical layer of the Semantic Web deems it imperative to incorporate inference rules into ontologies. Clark et al [117] have described OCL as the first order predicate logic that can be used to write constraints on object structures. It is this first-order inference rule that can be specified in designing ontologies. OCL is associated with both the model and the meta-model of UML. The complexity of learning a new syntactical language that is not easily implemented into an ontological paradigm makes it a difficult choice. But, for demonstrating purposes OCL is a powerful tool that should be looked into.

(b) F-logic can be mapped graphically using the Graphic Frame Logic (GFL). This can be represented visually and is extensible. It combines the features of UML with the dot code. It allows the mapping of F-logic into object-oriented programming and provides both XML and graphical representation [118].

7.3.2 Data modeling of neurological data sources

As seen in this work, the conceptual data models on physiotherapy and drug concordancy play a crucial part in untangling the ontological concepts and the relationships associated with them. It is imperative that to aid in the diagnosis of clinical research, the data models at both the conceptual and the physical layer be extended to house information on the dietary habits, speech therapy and occupational therapy. These databases can then be integrated and can also interact with each other by forming a federated database engine that will allow content to be displayed in traditional data format and machine enabled format like XML.

The implementation of the drug concordancy on the semantically enabled web was done for the purpose of data feedbacks from PWPD. With the gradual accumulation of legacy data, data warehouses consisting of data marts can be implemented to unearth functional segments of the clinical research, e.g. exploring the possible connection between socio-economic factors and drug concordance. Individual data marts can also be designed irrespective of their functionality so that it can be maintained locally while contributing to the distribution of information.

The heterogeneity of these data sources can also be extracted for the development of new concepts for ontologies. In this dissertation, one of the elements involved in the extraction of ontologies from RDBMS was reverse engineering. While it was possible to reverse engineer the data sources into ontological conceptual layers, the parent-child entity relationship in RDBMS did not get transferred into a subsumptional relationship between concepts. Since different reverse engineering tools provide different software knowledge representation, that is both syntactic and semantic, it is essential that research be conducted in the reverse engineering of databases to ontologies that identifies inconsequential concepts and removes the clutter.

7.3.3 Mobility of data

The transfer of data across different platforms and different geographical regions or across different time intervals was not practiced among the domain experts. The mobility of data would aid in the access of data at a time that is convenient to the users.

Replication enables data and database objects to be copied and modified from one database to another across different networks and platforms. Yet, the process of synchronization maintains the consistency of the database. The physical separation of the databases and latency are the integral part of the design process in replication. These characteristics enhance, among other things, the performance of the application. Other benefits [119], also include facilitating greater autonomy to users who can work with a local copy of the database and then transfer the changes to remote or mobile users across the network or over the Internet.

There are three kinds of replication. They are snapshot, transactional and merge. Snapshot makes a copy of the data and propagates the changes of the whole set of data rather than individual transactions, thereby making it a discontinuous process and entailing a higher degree of latency. Transaction replication allows incremental changes of data to be transferred either continuously or at specific time intervals. Merge replication permits a higher degree of autonomy and allows the subscribers to update changes and then propagates the changes to the publishers, which in turn transfers it to other subscribers.

7.3.4 Development of smart clients

The implementation of the PDCDB in the project was carried out using thin clients as discussed in chapter 3. Although in the project they were used for both external and internal users to input and access the database, the use of thin clients can be disadvantageous for mobile users. Mobile users can better run on smart devices like pocket PCs, Smart phones and other mobile devices. Data, as mentioned, in section 7.3.3 can be replicated and the services can be accessed using smart clients. They are typically used to provide access to essential data and services [116].

Smart clients make more efficient use of local and network resources, while providing support for occasionally connected users. Web services provide client device flexibility. Web services provide the building blocks for smart clients and the high interoperability of the web services allows it to be applied for a wider range of applications. Smart clients favour asynchronous communication, facilitate data caching and help in the access and mobility of data.