Chapter 2: Literature Survey

2.1 Introduction

It is essential to understand the interplay between the various disciplines that make up the workings of the Parkinson's Disease (PD). The different strands of neurological research like clinical, physiotherapeutic, dietary, 'speech and language' therapy store disparate and heterogeneous data that needs to be marshaled. Such conceptually different, yet correlated, data should be integrated to provide better information to the researcher. Maier et al [81] contented that since ontologies have the ability to provide a conceptual basis for communicating knowledge, the goal of integration is to consolidate the distributed data free from redundancy that will provide intelligent feedback and hence added values to both care givers and patients. While an N-tier architecture such as EJB (Enterprise Java Beans) in conjunction with network protocols can process the business logic and transfer data it does not provide the semantic sophistication of ontologies provided by languages such as RDFS [82] and hence does not have the capability to leverage the meaning of the data. It is this ability to provide a common vocabulary drawn from varied data sources that makes ontology an attractive paradigm for not only integrating the different data sources but also interacting between the user interface and the data sources. The use of ontologies in constructing a knowledge base and the integration of the heterogeneous data sources that help in the treatment of PD are the primary reasons for embarking on this research work.

As such the state of the art related to the project is reviewed in this chapter. It provides a description of the disease, the diagnosis and the treatment of patients with PD. A brief description of the available neurological diseases and the reason for the construction of clinical and physiotherapy databases is being provided. It elucidates the ontology design and model using the graphical representation of UML (Unified Modeling Language), construction of the knowledge base system and the two core constituencies of the knowledge layer of the architecture are described in section 2.8. This is followed by the ontological integration of the heterogeneous databases for the domain representation on PD.

2.2 Parkinson's Disease

No literature in PD starts without mentioning the seminal work of James Parkinson, entitled "An Essay on The Shaking Palsy" [15], where he for the first time identifies the cardinal features of PD based on his observations on six patients. He listed them as tremor, sleep disturbance, severe constipation, forward stoopness of the trunk, ***dysarthria*** and ***dysphagia.*** However, he did not mention muscular rigidity and facial immobility, which are also the essential components in the diagnosis of the disease.

2.2.1 Types of Parkinson's Disease

While Parkinson [15] attributed the pathogenesis of ***paralysis agitans*** to the lesions in the ***central nervous system***, it is now known that the lesion in the ***substantia nigra*** is responsible for PD [16], one among several lesions of the nerve centres. The other principal centres of localization are: (a) ***red nucleus***, (b) ***optic bed***, (c) ***Lenticular nucleus***, ganglion of the ***peduncular ansa***, nucleus of the ***vagus*** and the ***globus pallidus***, (d) ***Thalamic pallidus.*** Tretiakoff [16] examined the ***substantia nigra*** of nine patients and found the ***intracytoplasmic*** inclusions damaged. These damaged nerve cells are called ***Lewy bodies*** and are only detected after the death of patients. Although the clinical symptoms mentioned in chapter 1 is based on medical history and clinical examination, no biochemical tests exist to effectively diagnose the disease. This is complicated by the gradual progression of the degeneration of the Lewy bodies. This was noticed when 1 in 10 of 80 year olds were found to have incubated the degeneration of Lewy bodies without showing any signs of ***Idiopathic Parkinson's Disease*** (IPD) [17], the most common form of PD, which also means its cause is unknown. The onset of the disease after 50 years is known as late-onset PD while individuals with PD before the onset of 20 years of age is known as ***Juvenile-onset Parkinson's Disease.***

The characteristic symptoms of ***bradykinesia, rigidity*** and ***tremor*** are still prevalent in both the forms of PD. The genetic onset of the disease is manifested in (both the autosomal recessive and autosomal dominant genes) in the form of ***Non-Familial Parkinson's Disease*** and ***Familial Parkinson's Disease***. The genetic counseling of non-familial form of the disease in the family members has been found to be empiric

and the incidence of family members with chances of developing the disease is approximately 1-2 % [18]. The familial form of the disease is contained in a small number of families who are thought to have the disease due to the mutation of a single gene. Although six genes have been implicated, molecular genetic testing is clinically available only for PARK2, mutations in which result in the parkin type of ***Juvenile-onset Parkinson's Disease*** [8]. The other genes are SCNA, UCHL1, DJ-1, PARK3 and PARK8. α-synuclein, SCNA, is a small amino acid protein and has been found to be present in Lewy-like lesions for PD [19].

The absence of any clinical tests renders the diagnosis of the type of the disease difficult. Practitioners combining experience and clinical diagnosis expect PWPD to respond to anti-Parkinson drugs, such as levodopa. Such a 'challenge test' is used by doctors to measure the response of the patient to the drugs [20]. Hence, PWPD should respond to levodopa while failure to respond would cast doubts to the diagnosis. However, some drugs like anti-depressants, tranquillizers may lead to ***Drug Induced Parkinsonism*** and should be avoided by PWPD [21]. The Parkinson like symptoms disappears once the medication is stopped.

Patients who do not respond to the 'challenge test' but show some of the symptoms of IPD may be diagnosed with ***Progressive Supranuclear Palsy*** (PSP) or ***Parkinson-Plus***. Some of the symptoms of PSP include rigidity of muscles in the back of the neck and problems with vision, loss of balance and unexpected falls [22]. Like IPD, its cause is unknown and affects the ***neurofibrillary tangles*** unlike the IPD and affects 4-5 people out of 100,000 people in the UK [22].

Multiple System Atrophy (MSA) or ***Shy Drager Syndrome*** is characterized by symmetric parkinsonism without tremor and early postural instability and responds poorly to dopaminergic treatment and is caused by ***striatonigral degeneration*** [12]. Similar to MSA, ***Vascular Parkinsonism*** [12] is innured to dopaminergic treatment and does not exhibit any tremor but exhibits wide-based ***shuffling gait, dementia, pseudobulbar affect, urinary symptoms*** .

2.2.2 Diagnosis

The clinical diagnosis of the disease is made with a combination of patient history, medical examination and improvement of signs and symptoms with dopaminergic treatments [12]. The disease is differentiated from vascular parkinsonism, parkinsonism plus and drug-induced parkinsonism [23]. Increasingly, neuroimaging techniques are employed to assist doctors to differentiate between IPD and other forms of parkinsonism.

MRI brain scanning is used in the detection of vascular parkinsonism, IPD and drug-induced parkinsonism when the clinical findings are atypical [23]; Automatic Functional Tests or AFTs are used in MSA [20]. Vingerhoets et al [24] found that PD decreased striatal 6-[^{18}F]-fluoro-L-dopa (F-DOPA) uptake when measured by positron emission tomography (PET). Functional neuroimaging has been used to monitor the progress of the disease [24] and where clinical diagnosis is unclear, imaging of dopamine terminals with [β]-CIT is conducted [25]. Recently, a dopamine transporter chemical scan known as Dat Spect Scan or DaTSCAN has also been used to differentiate between IPD, drug-induced parkinsonism, MSA and PSP [20].

2.2.3 Medical Treatment

The two neurotransmitters, dopamine and acetylcholine work in harmony to maintain the balance and the coordinated movements of the body. In PWPD, the degeneration of the dopaminergic cells leads to the depletion of dopamine, which causes stiffness of muscle and impairs motor movements.

The purpose of drug treatment is to alleviate this imbalance [21] by restoring the balance between dopamine and acetylcholine. This is made possible by injecting drugs that replenish the dopamine content in the brain or by blocking acetylcholine. Since no cure is available to date for the disease, treatment lies in controlling the symptoms of the disease. Anticholinergics prescribed to younger people in the early stages of the disease have a propensity to improve tremor and can be used in the reduction of saliva and the damping of bladder contractions [21]. Dopamine agonists or levodopa is the preferential treatment for PD [12].

Dopamine agonists directly stimulate dopamine in the brain and are used in the early treatment of PD. However, the adverse effect of using it on elderly people warrants the usage along with levodopa. The efficacy of the dopamine agonists is considerably reduced due to the side effects of nausea, hallucination, confusion, leg ***oedema*** and dizziness related to low blood pressure [21].

The use of levodopa in the treatment of PD has already been mentioned in chapter 1. Levodopa or l-dopa is made in the body by tyrosine and the rate of formation is controlled by the enzyme called tyrosine hydroxylase that is then changed to dopamine [16]. It has been found that most patients treated with dopamine agonists will move to levodopa within five years [26]. The commonly available forms of levodopa are Sinemet and Madopar and contain an extra substance that prevents the levodopa being converted to dopamine before it reaches the brain. The advantages of levodopa are that it helps in the reduction of stiffness and slowness of movement. The disadvantages are similar to dopamine agonists except for the absence of leg ***oedema***. Long-term usage of the drug can lead to ***dyskinesia*** [27].

COMT Inhibitors break down levodopa by blocking the enzyme catechol-O-methyl transferase (COMT). This slows down the destruction of levodopa and hence it is used in conjunction with levodopa in order to prolong its effect. Increasing the duration of 'on' time can offset the 'on-off' switch prevalent in levodopa. It also helps to reduce the dosage and frequency of levodopa [21].

Before the advent of levodopa, surgery was the mainstay in the treatment of PD [12]. Currently, the following surgical techniques are employed in the treatment [28]:

- Lesioning (pallidotomy, thalamotomy and subthalamotomy)
- Gamma knife surgery
- Deep brain simulation (DBS) (thalamic, pallidal and subthalamic simulation)
- Brain implants using foetal brain tissue
- Infusion of chemicals in the basal ganglia (the part of the brain affected in PWPD)

The impact of surgery can lead to haemorrhage, seizures and death [29]. The side effects of DBS can cause dyskinesia, mood, speech and gait disturbances.

2.2.4 Living with PD

Due to the lack of any known cure, it is imperative that PWPD leads an active life. A prolonged campaign to confront this incurable disease will improve the quality of life. As such, a multi-disciplinary team approach for the managed care for PWPD is essential if the patients are to have a quality life [30]. Depression and cognitive changes are [31] common factors in the ailment of PD. Dietitians, Physiotherapists, Speech therapists along with Specialist Nurses, Consultants, Geriatrics and General Practitioners are an integral part of the multi-disciplinary team. The role of physiotherapists lies in the teaching of new skills or to help the patients relearn some of the old skill, key determinants in the physiotherapist's ability to improve the quality of the patient's life, while speech therapists help in resolving speech difficulties. Dietitians help in controlling the amount of protein when drugs like levodopa are taken while maintaining a healthy diet for PWPD.

The purpose of physiotherapy is "to maximize functional ability and minimize secondary complications through movement rehabilitation within a context of education and support for the whole person" [32] and deals with the core areas of gait, balance and posture [32], which are the basic activities of daily living. By early referral to physiotherapists, patients can benefit from the following: (a) address concerns about differential diagnosis, (b) assessment and monitoring early identification of movement problems, (c) introduce movement strategies for use over the course of the condition and (d) monitoring the drug efficacy to optimize motor performance [33]. This is made possible since physiotherapists can offer assessment, monitoring, treatment and management, referral to other services and providing support and information [34].

2.3 Neurological Databases

The need for a network of databases housing the clinical information would facilitate speedier progress of neuroscience only if the researchers shared their results in a

network of databases [35]. Researchers at Cornell University [36] have developed a repository for the neurophysiological recordings from neocortical neurons. Using a hierarchical classification, they designed a database that is general in its applicability. The SenseLab [37] integrates the olfactory receptor proteins with that of the electrical properties of proteins. Although these databases help in expanding the frontier of research in neurological sciences, none of them provides critical information to health care workers that help them to make the necessary decision that is vital to the healthcare needs of the PWPD.

In the pursuit of epidemiologic studies, Yesavage et al [38] using the US Department of Veteran Affairs pharmacy database for two sites evaluated the use of antiparkinsonian drugs like levodopa/carbidopa as an indirect measurement of PD for patients exposed to pesticides. They were then successfully able to find that patients at a particular site had a higher exposure to the drugs than at the other site.

Databases containing the structural formula, the activity of DA [39] and also the inhibitor sites of MAO inhibitors [40] have been developed. The surgical procedure of Deep Brain Simulation (DBS) involves the implantation of a wire with 4 electrodes at the tip into one of the three target sites of the brain: (a) the thalamus, (b) the globus pallidus and (c) the subthalamic nucleus [41]. The effect of DBS on the tremor of PD has been catalogued into a database that tests the Parkinsonian rest tremor velocity under four conditions of no stimulation and no medication, deep brain stimulation and no medication, no stimulation and 150% medication and deep brain stimulation and 150% medication [42].

It is in light of the absence of clinical databases regarding PWPD till date, that real time databases housing information relating to the drug concordance and its effect along with the physiotherapeutic aspect of the disease have been developed here.

2.4 Ontological Integration

The spectacular growth of biomedical and biological data scattered across the World Wide Web (WWW) in the recent past has ushered in an era in which the development of

the Semantic Web is being regarded as the key to harness the plethora of information available into tangible knowledge.

The advent of the web and markup languages like XML♦ has made it possible to reconcile disparate schemas of diverse data sources into an efficient and structured format that is interoperable [43]. The grand goal of the semantic web lies in collating the disparate data sources that will seamlessly merge and mediate with domain specific ontologies that can be used for different knowledge-based informational retrieval like query augmentation, ontological alignment, content aggregation or presentation [44]. An ontology is defined as a classification methodology for formalizing a subject's knowledge or belief system in a structured way (typically for consumption by a computer database). This means that the concepts and relations along with the axioms can be shared for a particular domain of knowledge. Knowledge engineers in confluence with domain experts develop ontologies by tapping into the tacit knowledge that is inherent in each of the experts. However, the software paradigm of ontology not only encompasses concepts, attributes and rules that constitute the schema but also entails instances or data.

Medical researchers, sifting through the desiderata of neurological data like that of PD require unambiguous communication to understand the workings of PD. For example, tacit knowledge of domain experts leads to different semantics of the same concept like ***Idiopathic Parkinson's Disease*** and ***Parkinson's Disease*** or ***Shy Drager Syndrome*** and ***Parkinson-Plus,*** which leads to potential misunderstandings during the process of making it explicit, as noted in our research. As such physicians, care givers and researchers are faced with the labyrinth of information to make a meaningful decision [45]. Knowledge relevant to clinicians for evidence-based clinical diagnosis, prognosis and treatment can only be explicated by reducing the "noise" of less relevant data. As such, the ontological approach of integrating the different data sources only aids in the enhancement of knowledge querying [46]. Researchers at the National Library of Medicine developed the UMLS, a taxonomy containing three knowledge sources, namely a Metathesauras that contains semantic information about concepts and relationships of clinical and biomedical data, a Semantic network that assigns all the

♦ http://www.w3c.org

concepts to a specific disease and a Specialist Lexicon that contains syntactical information about biomedical data. The absence of any formalism in UMLS does not make a powerful ontological alignment for our research purposes.

2.4.1 Ontological approach to Integration of databases

The pervasive nature of ontology ranges from controlled vocabulary (e.g. Gene Ontology♦) to full-blown ontology (e.g. PharmGKB♣). This dissertation advocates a template for ontology approach whereby data, metadata and semantics of diverse data sources can be integrated. The integration supports both the local and global views, and enables the conceptualizing of a framework, where data can be put in the context of analysis and knowledge extraction. The rationale behind this approach as elucidated is supported by literature [48] as follows: (a) the ontological commitment of data sources are minimal, (b) it preserves the richness of data sources and the flexibility of usage, (c) introduces no prejudice to data integration and (d) ensures semantic uniformity of heterogeneous data.

The structure of the biomedical ontologies can be categorized as 1, 2 and 3-dimensional. One-dimensional ontologies contain a set of hierarchical concepts without defining the relations between them (e.g. Gene Ontology) while two-dimensional ontologies have relationships between the hierarchical objects. The property of the object is dependent as it is tied to the relationship between the object and the value of the property. The 3-dimensional ontology has a hierarchical structure for objects and properties as well as a network of relations besides having a distinctive nature between the relationship of the objects and the properties. BAO (BACIIS Ontology) is a domain ontology for the Biological And Chemical Integration Information System (BACIIS) is an example for a 3-dimensional ontology. BACIIS employs the BAO ontology to map the ontology to each of the web database. This is made possible by mapping the data source schema layer to each of the databases [49]. The major advantage of BAO is that it makes no assumption about the format of the data sources. The schema mapping layer is then extended whenever a new data source is added instead of updating BAO.

♦ www.geneontology.org

♣ www.pharmgkb.org

Ontology as described in SEMEDA [50] is a pair:

O:=(N,E) …….(1)

Where N is a set of concepts and E is a set of edges, such that $E \in (a,b,t)$ where $a,b \in N$, and t is the type of relation which defines the semantic. Kohler et al [50], using the domain for molecular biology developed a 3-tier web-based system for "intelligent" semantic integration and querying of federated databases. One of the components of the system involved the implementation of ontology based semantic database integration where the database attributes were defined by referencing them to ontological concepts. The tables in the RDBMS are transferred as concepts while the columns of the tables are migrated as attributes or relations of the concepts in all the four databases. The mapping of the tables to the concepts of the ontology is possible due to the transitive nature of the 'is-a' [51] relationship. Concepts in ontology are referred to real-world entities that encompass the conceptual description of a domain while domain, ranges and cardinalites are treated as constraints.

The integration of data can be carried out in three stages namely, syntactic, semantic and after the verification and validation or curation of integrated data. In the case of syntactic integration, the fields are matched according to the names while the fields in the semantic integration are matched according to the semantic specification or domain concepts. As such, ontologies can act as a template in the integration of data, metadata and semantics of diverse data sources. While syntactic differences can be explicit and hard to reconcile, the semantic differences can be subtle and implicit. The differences in data models and data languages can be resolved in the syntactic heterogeneity while the differences in the underlying meanings of the representation of data [52], referred to as semantic heterogeneity, can be mitigated by the use of ontologies. Using the global scheme, Verschelde et al [52] integrated external and diverse data sources like Gene Ontology (GO) and Swiss-Prot databases to a proprietary biomedical ontology. In doing so they were able to improve on the expressive power of GO by proposing an improved representation on some of the 'part-of' relations like '***flagellum***', '***membrane***' and '***flagellar membrane***'.

Mitra and Wiederhold [53] developed the semi-automatic ONION (Ontology Composition) System whereby all external ontologies were first converted to the common format and the semantic heterogeneity among the objects was then resolved. The ONION system used the Graph-oriented conceptual model, expressed in RDF format, in the representation of the ontologies where ontology can be represented as follows:

$$O = (G, R) \quad \ldots\ldots\ldots(2)$$

Where G is a directed labeled graph and R is set of rules.

The graph G can then be represented as

$$G = (V,E) \quad \ldots\ldots\ldots\ldots(3)$$

Where V comprises a finite set of nodes and E defines the finite set of edges (or links) such that $E_{i,j} : (V_i , V_j)$. If the edges are ordered pairs then the graph is directed. On the contrary, if the edges are nonordered pairs the graph is undirected.

Barrett et al [54] using the RDF representation for the airline domain developed several infrastructure ontologies that described database structures, application services and query structures to support component interaction. This was made possible by mapping corresponding data sources to ontological concepts and relations. Relational databases (RDBMS) provide semantic structures or schemas and data or instances. Reverse engineering RDBMS into ontology facilitates the autonomy of databases by allowing semantic web applications to access the RDBMS via queries that exploit such integration. It also allows for the generation of semantic web enabled content (either in RDF or DAML+OIL format).

Different reverse engineering tools provide different software knowledge representation that is both syntactic and semantic [55]. While syntactic differences arising from the data types supported by each of the vendors can be resolved, semantic differences can be hard to reconcile since the information model captured by each of the tools is different [56] and the programming languages of each of the model representation can be different [57]. Jin et al [55] characterized conceptual transaction as the ability to extract implicit stored information in databases by identifying dependencies among

entities. They were therefore able to identify extracted concepts from reverse engineering as native, derived where the concepts need to be either derived or inferred from the facts represented, and undecided. In our case, the concepts extracted by reverse engineering were either native or derived. They used ontology restricted to the domain for reverse engineering tools in order to facilitate the semantic integration of the conceptual transaction adapter that exploits the full usage of such characteristic features of reverse engineering tools like architecture and graphical representation. But this method is restrictive only to those specific reverse engineering tools that participate in such integration.

In the case of semantic integration, there is no common language for describing and comparing data sources for which no standards currently exist, while in the case of metadata there is a shared repository for enumerating available data sources in machine-processable form while data can be stored as instances of ontology. However, with structural integration there is a common semi-structured data model using XML, and XML queries and transformations are used to resolve schema conflicts.

In the case of the global as view (GAV), the integrated database is described as a view either by the union or transformation of local databases, such that the queries executed can be translated to sources that need to be added with a concomitant change in the global schema. Gene Ontology, which has a broad-scope, employs such a view. The local as view (LAV) uses XML sources where each of the sources is described as a view of a virtual database.

2.5 UML Representation

Tim Berners-Lee envisioned the idea of the Semantic Web in which automated agents would scour for intelligent information, whereby software processes would interpret machine-understandable data linked to the web. The manifestation of the idea is intrinsically linked to the annotation of the web resources with metadata. Standard languages, like the RDF, RDFS and DAML+OIL [58], for such annotation are being developed under the auspices of World Wide Web Consortium♦ (W3C). However, automated processes would not be able to make use of the said languages unless there is

♦ www.w3c.org

a common vocabulary. It is in light of this context, that ontologies have come to be regarded as the building blocks of the semantic web that can be communicated across people and machines. While these languages have yet to formalise a standard graphical representation, Kogut et al [59] and Melnick [60] have leveraged the use of Unified Modeling Language (UML) [61] in expressing and modeling ontologies for knowledge representation.

A model has been described as "a simplification of reality" [62]. As such it can capture the essential features of the system by compartmentalizing the components that reduce clutter and hence focuses on those features that are the key ingredients to the system [63]. It is through modeling that we strive to understand the system while enabling us to better visualize, specify the structure and behaviour, provide a template for constructing, and finally create a document of the decisions that have been made for the system [62]. It helps both business analysts and knowledge engineers who interface with users to visually demonstrate the model while describing the domain. Modeling a knowledge system reduces the development costs since modeling involves the reiteration of the analysis and the design stage of the project before embarking on the developmental aspect of the project.

UML is a modeling graphical language, and the vocabulary, such as things, relationships and diagrams, and rules that it constitutes helps construct the artifacts of the conceptual and the physical representation of a system. Although graphical symbols enable us to visualize the model, it is made of well-defined semantics. This facilitates the interaction and ease of communication between a modeler and the developer since a model once designed can then be interpreted unambiguously.

Currently, no standard graphical tool exists to represent RDFS and DAML+OIL for the development of bio-ontology. The similarities between UML and DAML+OIL are many. Both have classes, which may or may not have instances. The generalisation/ specialisation of UML is translated as subClassOf in DAML+OIL. However, as one of the many differences between UML and DAML+OIL, the modularity of UML is not supported in DAML.

This dissertation proposes and discusses the use of UML Profile to facilitate the representation of a knowledge management system for PD.

2.5.1 UML as Knowledge Representation

The avalanche of neurological data available from disparate information sources warrants the use of knowledge management to provide meaningful information to medical practitioners. Knowledge has been classified as tacit and explicit. While explicit knowledge can be seen, shared and communicated with others [65], tacit knowledge has been described as things that are implied and not expressed openly [66]. The conversion of tacit to explicit knowledge is facilitated by the collaboration of knowledge workers in the knowledge acquisition phase. Knowledge management involves "methods and techniques for knowledge acquisition, modeling, representation and use of knowledge" [67]. Knowledge modeling has been used during the phase of knowledge acquisition so that it can structure, acquire, validate and store knowledge [68]. While frame based knowledge management tools like Protégé-2000 and Ontoedit have been used for modeling domain dependent ontology, the flexibility, ease of use and the unambiguous way by which UML can be used while tapping into the tacit resources of the domain experts is simply not available in either of the tools as the author found out during the course of the project. At the same time, the complexities of the model can be best understood when demonstrated visually since this enables the knowledge engineer to stimulate discussions among domain experts [65]. Protégé-2000 itself has a plug-in for UML to enable the modeling of ontologies.

The intuitive nature of the graphical representation of UML in modeling constructs has made it an essential tool in modeling knowledge based system. The Object Management Group♦by the use of its Model Driven Architecture♣ (MDA) is striving to make UML machine-processable, instead of its current usage as just a software modeling tool for the object-oriented programming paradigm. Cranfield [69] has enunciated the following reasons for the usage of UML in developing ontology:

(a) UML Class diagrams can be used to represent ontology

♦ http://www.omg.org/technology/uml
♣ http://www.omg.org/mda

(b) UML Object diagrams can be used as instances of classes and hence can be regarded as instances of knowledge
(c) use of OCL as ontological constraints
(d) the same paradigm can be used to model both ontologies and information systems

Class and Object diagrams that represent the static diagrams of UML are a manifestation of ontology in an object-oriented way. Class diagrams model the concepts, attributes, properties, and sub-properties, while the Object diagram is regarded as an instantiation of knowledge.

Taking a cue from modeling conceptual systems at different levels of abstraction in software engineering, Bergenti et al [70] modeled a multi-agent system using UML. By exploiting the notation of extensions, like stereotypes, in UML they modeled the basic agent-oriented concepts such as entity, agents, ontology and interaction protocol. The abstraction was conducted at the agent level and it was therefore possible to hide the implementation details, which are dealt within the object level. Ontologies were represented as classes of entities similar to the classes in the class diagram that represents a network of abstraction of objects. Attributes in UML correspond to the properties of ontologies, which unlike UML are first-class predicates. The attributes modeled the structure of the entities and their visibility is public. The relationship between the entities constitutes the metronomic aspects of ontology that map to the part-whole relationship of UML.

Using philosophical and psychological theories, Guizzardi et al [71] evaluated the well-founded ontological correctness to structured UML modeling constructs. They used the ontological primitives of classes, attributes, data types and associations to provide an interpretation of modeling a UML class diagram that can be used in conceptual and ontological modeling. An extension of this work was further carried out to develop a UML profile for Class [72] and Object types to provide an ontological representation.

UML has built in extensions that allow modeling primitives to be added geared to the specification of the domain. Such fine-grained specification in modeling constructs facilitates the establishment of the semantic differences that are essential in modeling ontologies. There are two different kinds of extensions provided by UML. These are:

(i) using the meta-model layer of UML to add new semantics and (ii) add new elements to the metamodel layer by changing the MOF♦ (Modeling Object Facility) model. The first extension is known as the lightweight extension while the latter is known as heavyweight extension. The basic UML profile consists of four modeling constructs of stereotypes, tagged values, constraints and tag definitions. This essentially means adding new modeling tools by extending the basic modeling tools. Stereotypes are enablers of virtual subclasses of metaclasses of UML. The purpose of using UML profiles is that it allows the development of new domain language based on MOF, which is inexpensive since it uses existing tools and it extends the UML Metamodel. A UML Profile provides add-on capabilities to ontological modeling like agent-based modeling [64] to connect both the information system and the artificial intelligence communities and can transform UML models into formal specifications that can be used to execute the system [73].

Such in-built extensions were used in introducing the benefits of UML to modeling biological systems and processes like enzymatic reaction [74]. Although it was modeled for the development of traditional information systems, they were able to delineate the different salient features in system biology by mapping to UML constructs (they called it SB-UML).

The integration of heterogeneous data sources and the corresponding interaction and maintenance of knowledge bases can give rise to knowledge acquisition bottleneck that can only be mitigated by the knowledge engineer with the help of a formal representation language that deals with the underlying knowledge system [75]. Configuring such knowledge bases can be achieved by the use of UML [62] and OCL, a graphical language that has wide found acceptance in the industry. Duric [76] using the UML profile based on the definition of Ontology Definition Model (ODM) developed the Ontological UML profile for identifying with the concepts of OWL, a language used in the development of the Semantic Web. ODM is still a work in progress that is trying to develop an ontological based language suitable for modeling ontology languages from the aspect of MDA.

♦ www.omg.org/mof

2.6 Knowledge management and biomedical ontologies

The modeling of the ontology is used in this dissertation for the construction of knowledge base systems. Ontology is regarded as the building blocks on which the semantic web is built. Ontology, as described by Gruber [77], is an explicit specification of a shared conceptualization. It is essential to understand what the phrase "explicit specification of a shared conceptualization" means before ontology can be used as a powerful paradigm in deciphering the semantic and syntactic differences that exist in the legacy systems of medical informatics. "Explicit specification" describes the concepts and the relations that exist between the concepts. Cogently, this permits the concepts to be arranged in a hierarchical manner. While "conceptualization" refers to the knowledge domain that is being addressed, "shared" refers to the confluence of tacit and explicit knowledge.

Explicit knowledge is represented in words, drawings, blueprints and equations that can be communicated to others [78]. Tacit knowledge is the knowledge that is accrued over the years by both individual experts and organizations. It is invisible and is difficult to communicate to others. It encompasses the knowledge that is acquired in the development of new skills, previous design sessions, possibly in the vendor's software tools or experience in the possession of key functional requirements that gives the competitive edge for the company. Since organisations change and people move either by promotion within an organisation or attrition out of an organisation and carry their tacit knowledge, it is therefore essential to convert this covert knowledge into explicit knowledge as quickly as possible for companies to retain the competitive advantage. Knowledge can include concepts, attributes, axioms and rules – the standard bearers of ontology. Ontology can, hence, be represented as a 5-tupel structure [79]:

$$O:=(C,R,H^C,rel,A) \quad \text{……..(4)}$$

Where C and R are identified as concept and relation identifiers,

H^C is a directed, concept hierarchy, such that $H^C(C1,C2)$ states that C1 is a subconcept of C2

The function *rel* is the relation which identifies concept as the domain and data types represent the range of the relation

A represents the set of rules or axioms, expressed in languages like F-logic

Knowledge engineers in confluence with domain experts help develop the domain ontology not only by tapping into the tacit knowledge of the experts but also by harnessing the plethora of resources of the institutions. The semantics of communication can be promoted by developing a common vocabulary, and hence the development of ontology for the specific business domain, that traverses the industry.

Maedche [79] defined knowledge base in the following 4-tuple format:

KB:=(O,I,inst,instr) ……..(5)

where KB is the knowledge base
O is the ontology as defined in equation (4)
I is the set of instances
the function *inst*, also called concept instantiation where $C \rightarrow 2^I$ and
the function *instr* also known as relation instantiation, is written as $R \rightarrow 2^{IxI}$

Ontology facilitates the conceptualization of the knowledge model and has become an integral part of biomedical informatics. Most of these ontologies developed are customised to their specific users. As such, their reuse and mapping to one other is a considerable hindrance in the development of KBS that has wider implications in medicine [28]. TAMBIS Ontology (TaO), based on Descriptive Language (DL), uses the conceptual framework of biological concepts and frameworks for proteins, motifs and similarities to access biological queries from data sources such as Swiss-Prot♣, CATH, PROSITE and BLAST databases. However, the integration of disparate data sources with ontology in the TAMBIS project remains a major research work.

A major aspect of the research is the integration of a database via ontology. The vast number of annotated databases, including mission critical, like diet, physiotherapy, speech and language therapy, and the drug treatment of patients♦ make it imperative to

♣ http://us.expasy.ch

♦ http://pdmp.cpe.surrey.ac.uk

seamlessly integrate these heterogeneous datasources that will provide the users the one-stop interface. Applications, although developed with the ability to interconnect with different datasources, do not necessarily eliminate the syntactic incompatibilities of data.

Gene Ontology consists of biological process, cellular component and molecular function to provide a controlled vocabulary that can be represented as a directed acyclic graph (DAG). Bodenreider et al [83] defines a controlled vocabulary as a set of terms that serves a common purpose to either index the literature review or annotate the terms while ontology is a framework of concepts and processes a system of hierarchical and associative relations that enables reasoning about knowledge. Their availability in XML format allows the incorporation in external databases. Gene Ontology has been incorporated in the Unified Medical Language System♦ (UMLS), a knowledge representation system that outlines both the broader and narrower concepts in correlating biomedical terms while the ontology developed by the Foundation Model of Anatomy (FMA) allows for the reuse of the ontologies by catering for the diverse audiences that require anatomical information [84].

2.6.1 Knowledge map

Knowledge acquisition is a hierarchical process. Knowledge is acquired incrementally [85]. Eliciting the intellectual capital that resides with domain experts and marshalling this into a knowledge-based system (KBS) format is a strategic imperative that most companies seek to gain a competitive edge. Knowledge captured can then be archived and shared among different levels of an organization, thereby encouraging reuse, preventing re-invention, saving retrieval time and acquisition costs [86].

Kim et al [87] using an industrial case study of a steel company built a knowledge map in which they outlined six steps of knowledge extraction, knowledge profiling, organizational knowledge, knowledge linking and knowledge map validation. Speel et al [88] defined knowledge mapping as the "techniques and tools for visualizing

♦ http://www.nlm.nih.gov/research/umls

knowledge and relationships in a clear form such that business-relevant features are clearly highlighted". They also developed the concept of the knowledge workshop whereby they captured and analyzed the organisational knowledge in a specific business domain and were also able to find out the existing gaps in their knowledge.

Formalisms were created for the development of a knowledge map of PD and the knowledge spine that acts as loosely coupled KBS within a federated structure. The aim was to develop a knowledge base for each of the specialist branches that a patient with PD encounters in the duration of the prognosis.

Visualization tools like UML were used as both personal and modeling constructs to enable the various experts conceptualize the domain and bring forth the tacit knowledge into an explicit one. The purpose is to demonstrate the development of KB on physiotherapy, speech and language therapy, dieting and drug-symptoms of patients with PD while enhancing the KB by drawing out inferred knowledge with the use of axioms. The benefits accrued in developing a KBS is not without constraints. Walczak [89] commented that the transient nature of knowledge, the cognitive differences that exist between experts and the knowledge engineers, the time taken to interview – one of the processes that was conducted -- are the major bottlenecks in the implementation and acquisition of knowledge. Knowledge mapping and merging were used instead of knowledge linking in the development of the knowledge map. The different domains of knowledge were then mapped manually to determine whether any particular element of knowledge is vital to one or more domains. This process is done following knowledge validation as this aids in the process of identifying any erroneous non-structural relations and duplication of concepts. Resources can then be queried from each KBS while each of the knowledge bases is then merged semi-automatically to become a central knowledge repository.

In order to define the concepts precisely, it is essential to outline the scope and context of the knowledge domain. Schulze-Kramer [90] categorized 3 kinds of domains, viz. (a) upper-level ontologies which outline the high level concepts within the framework of the domain (b) application ontologies which are based on the domain of the application and (c) task ontologies that are designed to solve a specific problem. Schweigert et al [91] have elucidated in detail the characteristics of biological data.

However, for brevity the following are listed: (a) complexity, (b) exception, (c) missing data, (d) changing models and data, (e) interoperability and (f) concept mismatch. It is in light of this context, that to overcome the semantic heterogeneity of different databases the usage of bio-ontology is noted. This rich structure enables information retrieval from various sources and hence facilitates the interoperability and communication between machines. The key ingredients in developing ontology lie in integrity, consistency and clarity [92]. Webster [92] has identified four new criteria in designing bio-ontology. They are granularity, abstraction, independence and isolation.

2.6.2 Mapping and merging of Knowledge Base

Mapping and merging of ontologies entail figuring out similarities that exist between the ontologies. Mapping allows the data structure of ontologies to be mapped. In a large knowledge base housing several thousand concepts and relations, manual mapping is tedious, time-consuming and prone to human errors. A human expert has to sort out any conflict that arises. Also, consistency checking and validation was not possible. Among the different tools surveyed, including OntoEdit, OilED♦, Protege2000♣, Chimaera♠, only Protégé 2000 and Chimaera supported merging of ontologies.

Lambrix and Edberg [93] using the REAL (Reliability, Efficiency, Attitude and Learnability) approach compared the user interface of Chimaera and the PROMPT plugin of Protégé 2000 for the task of merging Gene Ontology and Signal Ontology♦, an ontology for Cell Signaling System, and found that PROMPT provides a better overview of the ontologies, is easier to work with and the merging process is faster than Chimaera. They [93] also found that Chimaera provides a better functionality. The user interface was evaluated by testing with eight persons comprising four computer scientists and four biologists.

Developed by the Knowledge Systems Laboratory, Chimaera is an interactive tool based on the Ontolingua ontology editor that allows users to browse, edit and merge

♦ http://oiled.man.ac.uk

♣ protégé.stanford.edu

♠ www.ksl.stanford.edu/software/chimaera

♦ http://marine.ims.u-tokyo.ac.jp:8086/~spark/SO/

ontologies. It makes suggestions on two semantically identical terms based on the class-subclass or instance relationship, and allows the user the independence to make their own decision. PROMPT [94] is a Java based plug-in for Protégé 2000. It is not fully automatic but performs partial automation. It is an interactive ontology-merging tool and the algorithm is based on concept-representation structure, relations between concepts and the actions of the users. The algorithm generates a set for all the paths traversed and then calculates the similarity score for all the paths. It creates a list of suggested alternatives, provides possible conflict resolutions and logs knowledge-level ontology merging and editing operations.

2.7 Summary

The availability of biomedical data from different data sources makes it necessary to create new methods that will enable both clinicians and researchers to access information that is not only meaningful in their assessments but will also provide critical insights. The primary purpose of this chapter is to present an overview of the key research in the area of knowledge management that is related to neurological research, particularly in the field of PD.

In this chapter, the different neurological databases have been reviewed (section 2.3). The absence of any clinical database in the field of PD that helps researchers in improving the quality of life for PWPD is described in chapter 3. Databases, including annotated ones are created for the purpose.

The ontological integration of disparate data sources and the existence of neurological databases have also been reviewed. This includes the advantages and drawbacks of the integration of data at both the syntactic and the semantic level (section 2.4.1). The databases that have been created are then extracted for the creation of new concepts and the associated relationships between concepts. This is addressed in chapter 4.

Modeling knowledge bases using UML (section 2.5.1) has proved to be the viable technique due to the ease of use and the intuitive nature of the graphical language. However, the implementation of axioms in UML is an ongoing research topic due to its difficulty. This research focuses on the use of UML Profile in modeling a DAML+OIL enabled knowledge management system (chapter 5).

In this thesis, a novel method of modeling ontology that is captured for the construction of a new process of knowledge map in the development of the knowledge framework is presented. The process of knowledge map is used in the development of the knowledge bases that were then merged and mapped (sections 2.6.1 and 2.6.2). Both manual and semi-automatic mapping has been done in this work. Manual mapping suffers from the limitations of being tedious and labour intensive while semi-automatic mapping presented with challenges such as conflict resolution.

The framework used to describe the methods implemented in this research is presented in the following section.

2.8 Architecture

The architecture of the semantically enabled PD consists of 3 layers. They are as follows:
(a) the knowledge layer, (b) the storage or the database layer and (c) the Web or the application layer. This is illustrated in Figure 2.1. This research work concentrates on the knowledge layer and the database layer and the corresponding interactions between them.

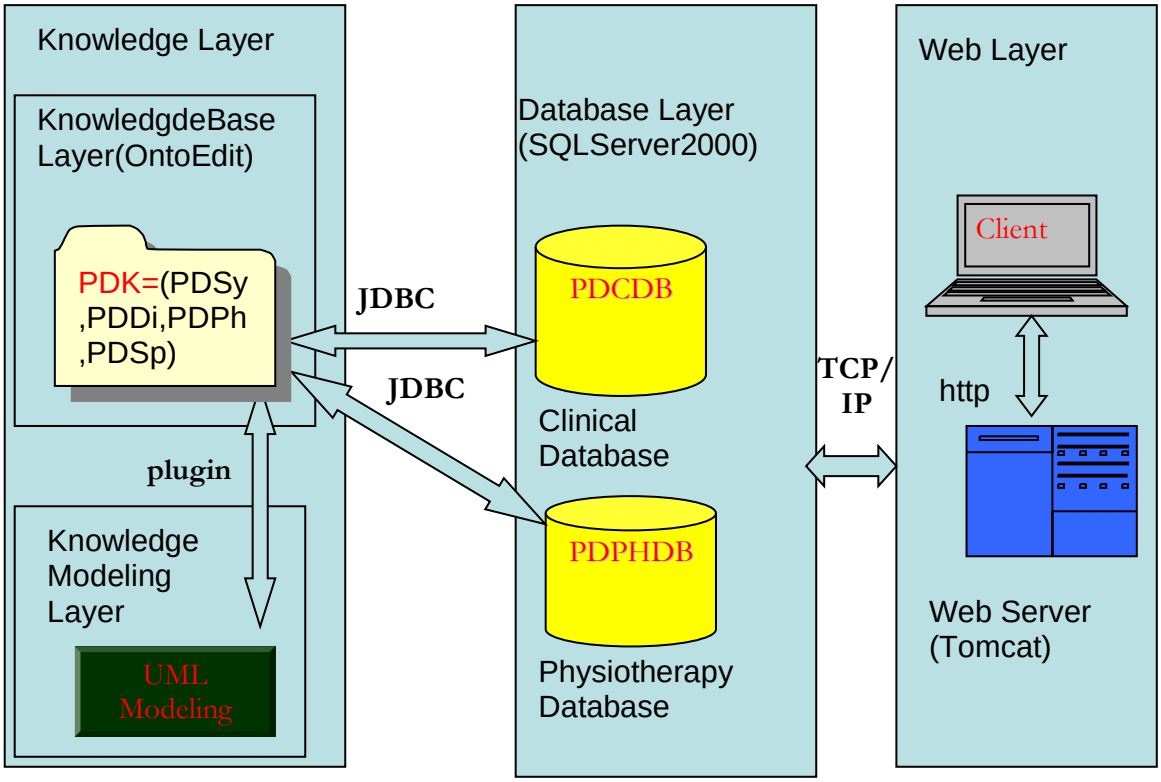

Figure 2.1 Semantically enabled architecture for the epidemiological study of PD

Knowledge Layer

The knowledge layer contains the knowledge base layer and the knowledge modeling layer.

The knowledge binding architecture provides a flexible infrastructure that bi-directionally interconnects knowledge to data sources which then connects data to the user interface. The knowledge layer consists of the knowledge base layer and the knowledge modeling layer. The **P**arkinson's **D**isease **K**nowledge base (PDK), as shown in the diagram, consisted of **P**arkinson's **D**isease **Sy**mptoms drugs (PDSy), **P**arkinson's **D**isease **Di**et (PDDi), **P**arkinson's **D**isease **Ph**ysiotherapy (PDPh) and **P**arkinson's **D**isease **Sp**eech therapy (PDSp) knowledge bases. Each of these knowledge bases was designed and developed in conjunction with domain experts to tap into the tacit knowledge resources and make it available in explicit form. The process of 'knowledge map' was introduced to develop the knowledge base. These knowledge bases were then

mapped and merged to reduce similar concepts and establish relationships between concepts across knowledge bases. The mapping and the merging of the different knowledge bases were conducted both manually and semi-automatically across the domain.

Client layer

The thin client layer was manifested in the form of the Web browser. This was done to eliminate the problems associated with the deployment and maintenance of an application that needs to be installed at different geographic locations for access to data from PWPD. This model also allowed the efficient exposure to a wider variety of external audiences within the targeted community. Many line-of-business applications are deployed using the thin client architecture as this facilitates access to both external users and those within the organization [2]. The business and the application logic for the semantically enabled portal reside on the server and the thin client makes processing requests to both the database and the web servers. The database server processes the query and then sends it to the Web server, which then fashions the result set in HTML format for displaying it on the browser. While latency is a disadvantage in thin client architecture, the requirement for the portal is simply to input the data from the user and store it in the database. As such, thin client architecture fulfils our objective.

A Semantic Web based portal on PD has been developed which is operational and is accessible♦. The purpose of the portal is to provide the latest knowledge to the community of doctors, nurses, therapists, patients, care givers and the general public about the causes and the impacts of the disease and its treatment. An important component in constructing the knowledge bank centres on the experiences of PWPD who live with the disease on a daily basis. The questionnaires are listed in 6 categories containing different aspects of the disease and most of them are in multiple-choice format for easy viewing by the users. The implementation of the thin client has been described elsewhere [3].

♦ http://pdmp.cpe.surrey.ac.uk

Database layer

The database layer consists of 2 databases, the **P**arkinson's **D**isease **C**linical **D**ata**b**ase (PDCDB) and the **P**arkinson's **D**isease **Ph**ysiotherapy **D**ata**b**ase (PDPHDB). Data gathered from PWPD through the thin client is stored in PDCDB. The web-enabled client transports data through the TCP/IP network layer to the database. Although, the PDCDB is designed and implemented for the MS SQLServer 2000, the existence of the conceptual model enables it to be migrated to any database vendor of choice. Queries have been constructed for both administration of the database and extracting information relevant to the purpose. The examples are shown in Appendix A. Since the conceptual model is symptomatic of meaningful domain representation, it is, therefore, possible to reverse engineer the semantics of the database model into ontology. The entities are translated as concepts while the attributes in the database are transferred as properties in ontology. The PDPHDB is an annotated database whose conceptual model has been reverse engineered to generate the concepts and the relations of the ontology.