CONSTRUCTING A BIO-HEALTH KNOWLEDGE BASE FOR ACCESS VIA A STANDARDISED ELECTRONIC HEALTH RECORD PROTOTYPE

Xia JING

School of Health Care Professions, Faculty of Health and Social Care, University of Salford, Salford, UK

Submitted in Partial Fulfilment of the Requirements of the Degree of Doctor of Philosophy, July 2009

Table of contents

List of figures

List of tables

Acknowledgements

I am hugely grateful to the Overseas Research Students Award Scheme and the University of Salford for funding my PhD study. Equal thanks go to the Institute for Health and Social Care Research in the University of Salford (MIE2005, HC2006, HC2007, HC 2008 and MEDINFO 2007), the Health Informatics Committee of the British Computer Society (MEDINFO 2007) and MIE 2005 for funding the conferences I have attended.

Millions of thanks are due to Prof. Stephen Kay, my great supervisor. Your broad knowledge, super understandability and intelligence, valuable guidance and suggestions, high professionalism and kind encouragement have supported me in getting through this hard, and painful, process of rebirth and growth. It will be too long a list if I write down here every bit of your help. I deeply appreciate all of it. Your academic and social merits will influence me beyond this PhD stage. Mr. Tom Marley, thanks for your intelligent guidance and cooperation, and your warm encouragement. To work with you enabled me to realise how amazing the human brain can be and made me feel I am living in the real world. Dr. Nicholas Hardiker, thanks for every of your helpful advice. Your sincere guidance, help and comments will never fade in my memory. Sincere thanks to Dr. James J. Cimino for valuable suggestions, helpful feedback and comments on the thesis. Deep thanks to Dr. Miao Sun for bio-genetics discussion and suggestions.

Thanks go to Mr. Ming Lu and Dr. Yongsheng Gao for endless and helpful discussion; thanks also to Mr. Andrew Schofield and Ms. Shijuan Li for discussion and information sharing; thanks to Mr. Matthew Horridge and Dr. Hai Wang for Protégé-OWL support; thanks Mrs. Su Ellis for your warm help; thanks to Mr. Ian Snape for proofreading Chapter 3 at an early stage. There are too many friends and administrative staff from the University of Salford to name here, many thanks for your warm friendship and individualised professional help throughout my study.

Thanks to my parents for bringing me to this fascinating world. Thanks to my Mum's standard which was set in my childhood and has become my drive through all these years; thanks to my Dad's compromise that always leaves me an exit in any road. Without my parents' expectation and encouragement, I could not have gone so far!

Declaration

No part of the work in this thesis has been submitted in support of an application for another degree or qualification of any university or institute.

The following publications have used part of the chapters in this thesis:

1. **Jing** X, Kay S, Hardiker N, Marley T. Ontology-based knowledge base model construction-OntoKBCF. MEDINFO 2007; 2007; Brisbane, Australia. 785-790 [part of Chapter 4]
2. **Jing** X, Kay S, Hardiker NR. Designing a Bio-health information assistant within an EHR. Healthcare Computing 2006; 2006; Harrogate England; 2006. 325-326 [part of Chapter 3]
3. **Jing** X, Kay S, Hardiker NR. Bio-health information: a preliminary review of on-line cystic fibrosis resources. MIE2005; 2005. 21-26 [part of Chapter 2]

Copyright

Copyright of this thesis rests with the author. Copies (by any process) either in full, or of extracts, can be made only for research purpose and must cite this thesis properly. Details may be obtained from the University of Salford Library.

The ownership of any intellectual property rights which may be described in this thesis are vested in the University of Salford. Any person intending to use the work described in this thesis for commercial purpose should consult Professor Stephen Kay at the University of Salford.

Dedication

For my Mum- Guanghua, Dad- Guangfeng and sister- Jian.

Abbreviations

CBO: Clinical Bioinformatics Ontology

CCR: Continuity of Care Record

CF: Cystic Fibrosis

CFMDB: Cystic Fibrosis Mutation Database

CFTR: Cystic Fibrosis Transmembrane Conductance Regulator

CIS: Clinical Information System

DB: Database

DDBJ: DNA Data Bank of Japan

EMBL-EBI: European Molecular Biology Laboratory-European Bioinformatics Institute

EHR: Electronic Health Record

EHRs: Electronic Health Records

EHR-s: Electronic Health Record System

EMR: Electronic Medical Record

GO: Gene Ontology

GP: General Practitioner

HIS: Health Information System

ICD: International Classification of Diseases

KB: Knowledge Base

LSDB: Locus-specific Database

MeSH: Medical Subject Headings

NCBI: National Center for Biotechnology Information

NHS: National Health Service

NIH: National Institutes of Health

NLM: National Library of Medicine

OMIM: Online Mendelian Inheritance in Man

OntoKBCF: Ontology-based Knowledge Base prototype about Cystic Fibrosis

OWL: Web Ontology Language

OWL-DL: Web Ontology language Description Logic

PDB: Protein Data Bank

PHR: Personal Health Record

RDF: Resource Description Framework

RDFS: Resource Description Framework Schema

SeRQL: Sesame RDF Query Language

SPARQL: SPARQL Protocol and RDF Query Language

UMLS: Unified Medical Language System

URI: Uniform Resource Identifier

W3C: The World Wide Web Consortium

XML: Extensible Markup Language

Abstract

Aim and Objectives: To explore the feasibility of accessing biological information and associated health information through a standards-based electronic health record. The objectives include constructing: a condition specific knowledge base prototype; an EHR system prototype based on a standard record architecture; and an interface that connects the two.

Method: An ontology was constructed to organise biological and health information in a formal and structured way. Cystic fibrosis was selected as an exemplar condition and the Continuity of Care Record was selected for an EHR prototype application. The sequence variations information and health information in the knowledge base are presented through the EHR prototype's interface and the results are evaluated.

Results: A substantive knowledge base prototype of cystic fibrosis was constructed. The content includes: the most common genetic mutations related to cystic fibrosis; time-oriented descriptions of cystic fibrosis; Cochrane conclusions; and gene therapy for cystic fibrosis. The content is organised on both time and problem oriented axes. It was found to be possible to present bio-health information that was case-specific through the EHR prototype interface.

Conclusion: Sequence variations information and associated health information can be made accessible through a standards-based electronic health record prototype. Complex knowledge can be accessed, to some extent automatically, thereby providing a starting point for integrating formal and structured biological information within health record systems which can be deployed in clinical settings.

Chapter 1 Introduction

1.1 Background and motivation
1.2 Research goal and aim
1.3 Research objectives
1.4 Research method introduction
1.5 Contributions
1.6 Structure of the thesis
1.7 Summary

In this chapter, the general research background and motivation, research questions, research goal, aim and objectives, major contributions and structure of this thesis are presented.

1.1 Background and motivation

There has been a huge and rapid increase in biological information as a result of the development of genetic technology. However, little of this information has been made readily available in the clinical setting[1, 2] Although biological information plays a critical role in understanding life activities, including health and disease status[3, 4], the connection between biological information and disease status has not been completely revealed and is still a matter for research. Another reason why such information is not readily available is that although there are many biological information resources for biomedical researchers[5], these have been designed or compiled with the researchers' requirements in mind rather than for the purpose of direct clinical use. Therefore they are not ideally suitable in a consultation scenario for example, in which there is limited time and where the participants are not necessarily specialists in genetics. Doctors in such settings do need biological information but only a selected amount, and they require it to be tailored to their use.

Here, the major research question relates to the feasibility of offering formal and structured biological information through systems likely to be found in the clinical setting.

Doctors already face information overload[6, 7] It is important that we do not make this worse and this is of concern in the selection of technology. Ontologies and knowledge bases are increasingly important tools in managing the complexity of knowledge and also have the potential for information reuse, sharing and extension[8](p2-5).

However, biological information is not an isolated resource and, consequently, providing such information will increase the quantity and complexity of information offered to the doctor and runs the risk of swamping the doctor. During the consultation between the doctor and patient a great deal of demographic and health information is already available to the doctor. As more becomes known in the domain knowledge area, the connections between the different types of information will increase dramatically.

Therefore it is desirable to use what doctors know as a way of interfacing their clinical knowledge with new biological information, and by so doing make the new information more accessible. Furthermore the amount of new information can be managed and controlled by using the patient's individual characteristics as recorded in their record.

In today's clinical consultation, the electronic health record (EHR) is becoming indispensable. It has become, in our view, not only a static repository for patient specific information but also an important dynamic tool through which to interface different types of information, including biological information, and a means of offering the doctor 'relevant' information.

A standards-based EHR was considered the best vehicle to use to carry out this research. Semantic interoperability has been identified as one of the most challenging parts for any EHR system[9, 10] and formal standards play a key part in solving such problems. Although this research only uses an EHR prototype, a standards-based prototype may benefit future standards development. This has the potential to influence many systems rather than a single isolated implementation.

In this thesis the concept of 'relevant information' refers to some definite relationship between an item of biological information and one or more items of health information. These relationships have been proven by biomedical researchers (domain researchers) and therefore provide a starting point for this investigation. These relationships are captured from published papers, text books etc in the domain knowledge ontology that will interface with the EHR.

The concept of 'relevant information' is also used within the EHR and an EHR system prototype had been developed to display such information. In the EHR system prototype general information is transformed to personalised information: i.e. the original set of information has been customised such that the information is shown as being relevant to a particular patient. [Note: in the individualised record, we may indicate not only confirmed relationships but also candidate relationships, i.e. factors, which the doctor may wish to consider or have investigated.]

1.2 Research goal and aim

The ultimate research goal is to bring together the different types of knowledge seamlessly, in order to illuminate disease status with respect to genetic information. A more modest research aim for this PhD is to explore the feasibility of offering formal, structured bio-health information that is relevant to health care in a simulated clinical system.

1.3 Research objectives

The research aim comprises three objectives and these can be considered as a set of paired abstract and physical artefacts. This research therefore uses and constructs specifications, which are at the abstract level, and maps these to physical artefacts which are developed on the concrete level. The objectives are:

- The organisation of biological and health information in a formal and structured way; at the abstract level this specification is a schema or model for a knowledge base; at the concrete level, a knowledge base prototype is constructed.
- To simulate a clinical system to access and manage bio-health information. At the abstract level, a standard specification for an EHR is used; at the concrete level a standards-based EHR prototype is developed.
- To interface biological information with health information so that both are accessible. At the abstract level this is specification of bridging points common to both; at the concrete level this is a set of headings and containers that show the content and structure of the knowledge base prototype, via patient specific filters.

1.4 Research method introduction

For our purposes cystic fibrosis was selected as an exemplar condition; the content was formally structured into the knowledge base prototype so that it might be used to deliver case-specific bio-health information. The formal and structured biological and health information was organised by a knowledge management editor: Protégé-OWL[11] An EHR system prototype, which was built using Microsoft Visual Basic and Microsoft Access, used the international standard Continuity of Care Record (CCR)[12] The knowledge base prototype was connected with the system prototype. The personalised biological and relevant health information has been made accessible through the EHR interface and has been tested by hypothetical test records.

1.5 Contributions

There are several contributions claimed from this research project:

- The bringing together of sequence variations information with health information using a standards-based EHR.
- The construction of an ontology-based cystic fibrosis bio-health model (the first of its kind) and knowledge base prototype.

1.6 Structure of the thesis

The major structure of the thesis is as follows:

Chapter 1: The Introduction chapter discusses the thesis in a general way. It includes background, motivation, research questions, research goal, aim and objectives, major contributions and structure of the thesis.

Chapter 2: The Literature review chapter introduces key topics and provides evidence related to:

- The motivation behind this research and the reasons why it is important to integrate biological and health information within the clinical setting.
- Current examination of existing research in similar fields and a comparison with this research.
- Background introduction to semantic web technology; the reasons for the selection of semantic web technology to represent and organise the relationships between biological and health information.
- Introduction to the EHR.
- Introduction of existing biological information resources and tools.
- The choice of cystic fibrosis as an exemplar condition.

Chapter 3: The Methodology chapter introduces:

- The research conceptual framework and research strategy.
- Detailed solutions to achieve the research objectives and their evaluation.

- The requirements and functions of the solutions.
- The reason for choosing CCR and the development environment of the prototype.

Chapter 4: The knowledge base prototype- OntoKBCF construction chapter includes:

- How the ontology-based knowledge base prototype is constructed.
- The decisions underpinning the scope, granularity, organisation and construction of the knowledge base prototype.
- The reasons underlying these decisions.
- The construction result and its limitations.

Chapter 5: The EHR system prototype construction and connection with OntoKBCF chapter includes:

- The overall design of the CCR-based EHR system prototype.
- The interfacing and integration of the bio-health ontology with the EHR system prototype.
- Testing of customised bio-health information through the EHR system prototype by test records.

Chapter 6: The Discussion chapter introduces:

- The research significance and implications.
- Research limitations and some technical obstacles.
- Potential benefit
- Future directions.

Chapter 7: The Conclusion summarises the research and presents recommendation for future work and the conclusion of the thesis.

1.7 Summary

In this introductory chapter the general research context and the research questions have been introduced and defined briefly: what are the questions to be focused on and why this research question is important. The research goal, aim and objectives have been listed. What contributions have been claimed based on this research and the structure of the thesis has also been included. The next chapter will review the literature and consider the existing work. The key topics will be discussed and the evidence for taking this approach will be set in context.

Chapter 2 Literature review

In this review chapter the major topics dealt with are: the research context and motivations, research that is related to the study; and the rationale for choosing the particular solution; introduction of technologies and the exemplar condition used in this research.

2.1 Terminology

It is necessary to clarify what is meant by 'biological information' in this thesis, because it is a very broad term. In this thesis 'biological information' is intended to refer to genetic information in general, and more specifically still to 'mutation information'. 'Mutation information' is still a very general term and actually 'sequence variation' is used as a starting point. The detail of the exact sequence variations of interest in this thesis will be discussed in Chapter 4 (4.3). The genotype is a type of biological information. The genotype is the genetic constitution (the genome) of a cell, an individual or an organism. It can pertain to all the genes or to a specific gene[13].

Health information is also a very broad category, covering many concepts. In this thesis it is used to refer to the organism level description which is restricted to the human being; for example description relates to diseases, symptoms, diagnostic or treatment related information and demographic information. Health information is typical of the information which is recorded in the electronic health record. The phenotype is the appearance or physical manifestation of an individual, which results from the interaction of the person's genetic makeup with his or her environment[13] There are three levels of phenotype: molecular, cellular and organismal[14] In this thesis the organism level disease phenotype is referred to as a type of health information.

2.2 Research context

2.2.1 Bio-health informatics research

With more and more biological information being discovered, researchers are trying to cooperate to bring biological information together with health information, and to explore the merging of the two areas (bio-health informatics, BHI)[2, 15] Martin-Sanchez and colleagues[2] named the new field as biomedical informatics (BMI). The merging and cooperation will bring benefits for both biological research and health care; the benefits for biological research include more precise mechanisms in interpreting life phenomena and disease status etc; the benefits for health care include new diagnosis and treatment methods, personalised health care service etc[2] Blake[16] also held a similar view about biomedical research: "genetic elements played a fundamental role in understanding disease, and information systems were major tools for data management"

Bio-health informatics research faces many challenges, of which some of these are of direct concern to this research:

- To examine genetic information in more than one health care situation during the entire individual's life; to offer correct interpretation of the genetic information to the health care professionals during the consultation[17].
- To include sequence and genetic information within the medical record as more new biological information is revealed[18] (p640); to include gene information in the individual's medical record will be a requirement for future molecular medicine[4]
- To present the sequence information to clinicians in useable and useful ways within the record[18]

It is still a major challenge to identify the genetic contributions to disease, i.e. to elucidate the precise relationships between biological information and health information, especially between genotype and phenotype[3], and there have been many explorations of this topic. In addition to experimental validation of the phenotype through the genotype[19], other research attempts: to analyze the relationship between genotype and phenotype based on accumulation data from both genotype and phenotype related to particular genes- for example WT1 and CARD15[20, 21]; to infer genotype from the clinical phenotype through knowledge-based algorithms for simple genetic diseases[22]; to develop a knowledge framework for cancer's molecular-disease relationships, which is a foundation for further automatic computation[23]; to analyze possible relationship between clinical phenotype and genetic information by clustering phenotype[24]; to develop visualisation tools to display genotype-phenotype relations from OMIM[25]; to bridge genomic with clinical discovery science by using knowledge base of clinical decision support system for genomic discovery[26]; to use natural language processing to extract phenotype information from literature[27]; to use mathematical methods to relate survival times with gene expression profiles[28].

Although there is much research into biological information and health information, it is hard to identify the precise relationships. Often the research is based upon a model organism instead of the human being. And it is very hard to conduct research starting from the organism phenotype, because the organism phenotype is hard to standardise and quantify. However, if there is sufficient genotype and organism phenotype information available, there will be great potential in establishing meaningful relationships between them. There are already huge amounts of organism phenotypes in the clinical setting, especially in clinical records. Consequently if the genotype information can be integrated into the clinical records then the clinical record itself becomes a rich environment in which the research on the relationship between genotype and phenotype may be conducted.

This research provides biological information into a clinical software environment, and seeks matching points to integrate biological and health information. This is an extension and a complement to current BHI research. Considering the important potential that genetic information might play in future medicine (detail will be in 2.4.2), bringing biological information into the clinical setting is a worthwhile pursuit.

The following is a current review of the two research directions that are related to the solution of this research: i.e., to provide formal and structured bio-health information of cystic fibrosis (i.e. exemplar) in a knowledge base through a simulated clinical system EHR system prototype.

2.2.2 The absence of biological information in the EHR

As noted, with the increase of biological information, many scientists and researchers have identified the need to bring biological information into the EHR system[29, 30]. However, there are at present few concrete published examples of research on this topic and most papers are either from a subjective viewpoint offering insight and/or prediction. Authors of the published papers have tended to focus more on what should be done rather than what has been done in this particular area although they do usually agree that biological information plays a key role in the clinical setting.

Hoffman[29] suggested developing a new laboratory system to include formatted molecular diagnosis and cytogenetic findings and to adopt a new standard/vocabulary for clinically significant genomic information. The view given involves including genomic information into an EHR system but without stating what type of genomic information should be included or how to achieve this view.

Sax and Schmidt[30] reviewed the available standards' formats for phenotype and genotype data, and assessed the data model for adequacy of representation. However the comparison and assessment given were quite general, without any detailed explanation being given about how or what to use in a real situation. They were more concerned with identifying current problems in this area rather than addressing how to solve these problems.

Robson and Mushlin[31] studied how to assemble patients' clinical genomic records, especially for patients' protein information in clinical records. However, the authors' focused on the technological aspects of transmission of clinical and biomedical data. Clinical data was formatted by the clinical document architecture (CDA) while genomic data consisted of the DNA sequence and protein sequence. However, in that paper Robson did not explain how to make the annotation data formal or structured, nor how to use genomic data in a clinical setting, but focused instead on encoding and decoding.

In the HL7 Version 3 set of standards there is work related to clinical genomics, i.e. Clinical Genomics, Pedigree which is under the clinical domains[32] This is a specification for communication of a patient's family history between different parties, such as healthcare professionals, patients, genetic test suppliers and EHR systems. The genomics data is a part of the family history. The interoperability is the focus of the specification. However, although the same reference model for the EHR communication is used in the CDA, little progress has been made to co-represent the genomic and health data.

Current research into trying to include biological information in EHR is still immature and incomplete. The ambiguous relationship between biological information and health information makes it more difficult to decide what type of data (gene data, protein data, structure data etc) should be included and how detailed these data should be. Is it feasible to include formal and structured biological information associated with health information in an EHR system? If so, how should biological information be included? These are two problems which still lack clear and definite answers from the literature. Several related but different research projects are introduced and are later compared with the work in this research.

2.2.3 Current research on biomedical knowledge bases

There are several reviews which discuss the trends and challenges in biomedical ontologies: for example, ontology was considered as a method to consistently annotate features from genotype to phenotype for biologists[33]; there are desiderata for reference ontologies in biomedicine, and five desirable characteristics for reference ontologies are presented e.g. "good lexical coverage, good coverage in terms of relations, compatibility with standards, modularity, and ability to represent variation in reality"[34]; the principal and the current issues in designing and developing bio-ontologies: the ability to query across databases and the constructing ontologies that describe complex knowledge (such as phenotypes)[34] These are valuable principles/suggestions for knowledge base construction.

Rubin and colleagues reviewed the research which attempts to offer both biological and health information from the same knowledge resource[35]. On investigation, it became clear that there are some defects or missing parts in many of the projects reviewed; for example the format is not formal or the content is not structured or health information especially the phenotype is simplified with only the disease name. To the author's knowledge none of the existing knowledge sources that were reviewed focus on both genotype and phenotype information which is for direct clinical use. Every project has its own purpose and associated advantages and disadvantages, but for brevity the following introduction will focus more on the missing parts.

HGVbase is a phenotype/genotype database at a summary level (group level aggregated data) instead of at the individual level[36] and HGVbase describes relationships between human DNA variations and phenotypes. There were around 119 different studies (diseases) in the new version of HGVbaseG2P[37], however, for most diseases the phenotype was only disease name without further details; the phenotype is much more complex than a single disease name. HGVbase was based on conventional database technology.

PhenomicDB is a new cross-species genotype/phenotype resource developed in Germany[38] and it shows phenotypes associated with corresponding genes and grouped by gene across different species. Although there were phenotype description, phenotype keyword and phenotype ontology ID in PhenomicDB, only the phenotype description had more detail, although this was in free text.

PharmGKB[39, 40] is a pharmacogenetics and pharmocogenomic knowledge base which includes gene, drug, disease and effects of gene variation on their relationships; knowledge resources come from both the literature and from original experimental data. Phenotype data follows certain formats and terminology in PharmGKB, but with more molecular and cellular phenotype information; clinical measures were content related to lab results and pharmacodynamics data.

Other related work include Gupta's Parkinson's disease ontology, which was designed for inference and computation[41], toxicogenomics[42], and asthma[43] knowledge resources, both of which were based on database technology.

From the above review it seems that the relationship between genotype and phenotype, the phenotype representation, ontology-based data storage, and the management and interchange of these data types are attracting attention to knowledge bases in biomedicine.

2.3 Comparison of this work with other related research

2.3.1 Delivering relevant information in the clinical setting

'Infobuttons' is a project led by Dr. James J. Cimino[44, 45] at Columbia University. It is a study to provide automated, context-specific links between the clinical information system and other online knowledge resources. Infobuttons is a very practical means of helping clinicians get relevant information within the clinical information system (CIS). The information needs came from observation of clinicians, surveys, focus groups interviews, and analysis of system log files, scenario-based assessment and knowledge acquisition[46-49]. The results provided by Infobuttons are links, and this strategy removes the need to update remote knowledge resources. Using generic questions is a creative solution for identifying retrieval questions. Infobutton Manager is used to create generic questions for the query. Other contextual information including user type, patient age and sex, the clinician's task etc is also used to filter query results. This is an effective way to tackle information overload in the process of information query within clinical system. However, Infobuttons was limited to handling specific questions[44] Generic questions sacrifice query precisions. However, if every single specific question was listed usability would suffer. This is seen as a difficult balance point among query precision, usability and information overload.

Both Infobuttons and this research try to 'link' outside information resources with the clinical system to provide relevant information support. The information resource in this research focuses on bio-health related material, which has an overlap with Infobuttons' However, it is not clear how Infobuttons handles biological information during the information support, whereas a major aim of this research is to bring biological information to the clinical system, especially sequence variations information. Nevertheless, this is a long term goal that will require extensive large scale testing beyond the scope of the present work.

Sheth's[50] work was an application of semantic web technology in the electronic medical record. However, the work focused more on the management side, especially billing and insurance checks and cover. On the health care side the work focused more on medication, for example guidance on dosage, interaction check and allergy reminder etc. OWL ontology and semantic rules were used in representing domain knowledge and allowed the system to make decisions. The paper did not mention if or how biological information was dealt with.

2.3.2 Delivering biological information in the clinical setting

Hoffman's[51] research aimed "to develop a curated resource, the Clinical Bioinformatics Ontology (CBO), a semantic network appropriate for describing clinically significant genomics concepts." His research combined the medical vocabulary with biological concepts which can be used in molecular finding: mutations, polymorphisms, allele naming and chromosome structures. Hoffman also imported the resource into a commercial healthcare information system (HIS). His work provided term/vocabulary support for the HIS. Hoffman's research and this research share the idea about the technology choice- both projects are ontology-based, and the general target fields biological and clinical fields, but the application and functional intention of his research are different.

Kumar's[52] research of connecting proteins to biological processes and establishing relationships not represented in Gene Ontology (GO), aimed to provide a unified representation which could help in answering the difficult questions which arise at the borders of medical informatics and bioinformatics. However, Kumar did not define the scope of the work in that paper. Theoretically, the wider the scope of a knowledge base the better. However, it is neither realistic nor practical to complete if there is not a clear boundary. The involved relationships in the ontology had been explained in detail by logic. Both statistical and probabilistic approaches were used to find associations among GO terms. There was no description of how the associations and the concepts from disease level to molecular level were organised.

PheGe[53] was a platform for exploring genotype-phenotype relations at cellular and organism levels, which was aimed at automatic network analysis on the organism level. This was a project based on database technology and a pilot scheme for virtual system biology. PheGe assisted in analyzing regulatory, metabolic and neuronal circuits; this is totally different from the aim of this research. PheGe and this research share the initial aim of helping to reveal relationships between genotype and phenotype; however the objectives and methods are different.

Although generally all the three studies have attempted to bring biological information into a clinical environment, they are quite different from this research either in approach, explicit granularity or in the type of technology used. Hoffman's research had been focused on biological vocabulary support, whereas this research stresses knowledge fact support. Kumar's research lacked detailed description about how to organize the concepts from disease level to molecular level, whereas this research uses this as a main part of the integration. PheGe research was based on database technology, whereas a key part of this research focuses on knowledge base technologies as a way of managing the complex semantics.

2.3.3 Other bio-health resources

Genetics Home Reference[54, 55] is a consumer resource designed by researchers from the University of Missouri and NLM. It is a web site about genetic conditions and the genes or chromosomes responsible for those conditions. Most of the contents are listed as questions and answers with cross reference in free text; it is user-friendly. This research focuses upon clinical use, hence the contents are narrower, the knowledge representation is more formal and the target user is a non-specialist doctor, such as a GP.

Gene Ontology (GO)[56] and Unified Medical Language System (UMLS)[57] are tools for terminology (vocabulary). Again, such tools are very useful but would only contribute part of a solution based on the requirements of this research.

Although these are similar in some respects, this research aims to merge biological and health information in a formal and structured way and to make the information accessible through a simulated EHR system which makes it quite distinct.

2.4 Research motivations

The motivations for this research are summarised:

1) There has been a huge and rapid increase in biological information as a result of genetic research, however little published research has focused on to making this information readily available in a clinical setting.
2) Biological information is one of the most important material bases for the understanding of life activities, including disease and health status. However, the connection between biological information and disease status is still being studied.
3) There are a large numbers of sequence information resources already in existence, but they have not been designed and compiled for direct clinical care. Therefore they are not ideally suited to a consultation scenario, in which there is limited time and where the participants are not specialist in genetics.

2.4.1 The growth and use of biological information

In the past decade with establishment of new technologies (e.g. automation of DNA sequencing and large-scale sequencing) and the completion of Human Genome Project, there is more and more biological information available, especially with respect to gene sequences. In GenBank, the genetic sequence database in the US National Institute of Health, an annotated collection of all publicly available DNA sequences, the base count rose from 9.5 billion in Aug. 2000 to 85 billion in Feb. 2008[58, 59]. According to statistics from the Protein Data Bank (PDB) web sites, in PDB the number of protein structures which can be viewed were 5008 in 1996 and rose to 51079 in 2008[60].

Despite the increased pace of identification of gene sequence and protein structure, biological information remains poorly understood and used in clinical practice[3, 61] (some examples will be given in section 2.4.3). Because the precise relationship between biological information and diseases phenotype has not been elucidated explicitly, it is hard for doctors to use biological information directly in their daily practices.

For biomedical researchers (domain researchers), the next step is to translate biological sequence information into health benefits[3], such as more effective treatment or faster and reliable diagnostic method. The intelligent and proper use of human gene sequence data could lead to significant advances in our understanding of genetic basis of diseases at the molecular level, and further tangible health benefits[61].

Although many biologists and biomedical researchers predict the critical role and bright prospects that genetic information will play in the medicine of 'tomorrow'[3, 17], there is a lack of answers to some basic questions, for example 1) how to use genetic information in clinical setting, especially during consultation; 2) what to use? At this time it would be true to say that most research is exploratory.

2.4.2 The role of genetic information

An important part of the argument for this research is the increasing importance and pervasiveness of biological information, as described in the last section. In this section the focus will be why genetic information is important. In this thesis 'genetic information' is used to refer to nucleic acids and proteins information specifically.

Nucleic acid is the material basis of life. A nucleic acid is composed of chains of nucleotides. A gene's nucleotide sequence encodes the amino acid chain, which coils and folds to form the functional protein[62](p174-177). Figure 2-1 is a representation of the biological hierarchy. It is the nucleic acid that guides a protein's synthesis, and it is the proteins that are functional for life activities, for example structural support, protection, transport, catalysis, defence, regulation and movement[63] (p42). So the nucleotide sequence provides a key and basis in determining the organism. There is an example of how protein affects phenotype in cystic fibrosis in section 2.8.3.

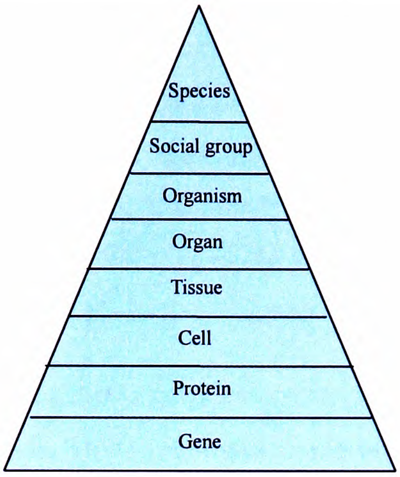


Figure 2- 1 Diagram of basic biological hierarchy

The ability to produce the correct proteins depends on nucleic acid, or more precisely, upon the DNA sequence[64](p46). "Abnormalities in nearly all classes of proteins, including enzymes, transport proteins, receptor proteins, and structural proteins, have been implicated in genetic diseases"[63] (p397). For example: an abnormal single enzyme, which is dysfunctional - phenylalanine hydroxylase- is the cause of phenylketonuria (PKU), while abnormal haemoglobin is the reason for sickle-cell disease[63](p377-378). There are also environmental factors involved with the disease status apart from the DNA sequence: "for most common diseases it is interactions of many genes and proteins and the enviromnet"[63] (p379).

There are about 20000-25000 genes for human DNA. It is possible almost all human genes can cause disease if they are altered substantially[65, 66]. In the following context haemophilia was used to describe possible relationships between genotype and phenotype in genetic disease. The genetic material basis for cystic fibrosis will be introduced in section 2.8.3.

There is often a logical relation between a person's genotype (DNA sequence) and their health outcome (phenotype). For example, there are quite different clinical presentations for haemophilia patients: patients with moderate bleeding problems can live to adulthood and lead relative normal lives without trauma or surgery; whereas a patient with severe spontaneous bleeding in their early childhood will rarely survive to adulthood. There are also patients in between with intermediate clinic severity[67]. The various mutations of the haemophilia A gene correspond to different health outcomes: a large gene deletion in the DNA sequence of the haemophilia A gene results in the failure of gene transcription, which then leads to severe status (the patient has severe spontaneous bleeding and rarely survives to adulthood). Genetic inversion and a small change in DNA sequence of the gene leads to amino acid substitutions in factor VIII protein, and causes mild to moderate disease[67]. The various mutations explain the different phenotypes in haemophilia A, however the correlations between genotype and phenotype in monogenetic diseases such as cystic fibrosis, are not always linear. The relationship between genotype and phenotype is much more complex.

Genetic information plays a critical role in understanding genetic diseases' mechanisms, especially monogenetic diseases. This is the main reason to choose genetic information, as it is a key material basis for life phenomena in both health and disease states.

2.4.3 Clinical use of sequence information resources

With the huge increase in the amount of biological information, especially sequence information, more and more sequence information resources are becoming available, for example: Nucleotide[68], DDBJ[69], EMBL[70] etc. It has been noted, however, that most of these resources have not been designed for direct clinical use, nor follow the clinical related recommendation[71]

Besides these general sequence information resources, there are also some locus-specific databases (LSDB, i.e. gene-specific mutation database)[72, 73], some of which involve both genotype and phenotype. The LSDB with phenotype has the potential to be used clinically. For example NPC-db[74] was a disease database of Niemann-Pick type C disease, which is an autosomal-recessive disease. NPC-db included both variations information and phenotype data. RAMEDIS[75] was a metabolic diseases database, which included variation information and structured clinical information. However, in NPC-db, the author did not mention how the phenotype data was organised and it is not clear if the phenotype data was free text, terminology compatible or structured data. In addition both NPC-db and RAMEDIS are based on database technology and have not been integrated with any electronic health record.

2.4.4 Information overload

Information overload is a big challenge in the information era and is a fact faced by doctors. Information overload comes not only from increasing information amounts, but also from the diversity of available information[6]. More than 400000 articles are added to the biomedical literature each year[7]. In addition to published articles, other information resources include textbooks, guidelines, references, databases, knowledge bases, discussion groups, newsletters, e-mail alerts, internet sources etc. When the increasingly amounts of health information are joined by rapid increases of biological information, doctors' information overload, already acute, becomes critical[6, 76] Offering biological information in the clinical setting should be designed to not make the information overload situation faced by doctors even worse.

Different methods have been used to help alleviate the information overload, such as offering a guideline summary by an official group[7]; offering a web-based intranet application[77], or developing a knowledge management system[78]. The emergence of semantic web technologies, however, brings new hope in data sharing and reusing, both of which are critical in handling information and the associated diversity. Semantic web technology provides some relief for information overload from the file format perspective, while the ontology-based organisation of knowledge facts provides some relief for information overload from the file content perspective. A detailed introduction to the semantic web and ontology will be given in section 2.5.

Reference sources of new information and knowledge have both advantages and disadvantages for busy doctors. From an ethical perspective, no one would deny that evidence based practice is the ideal; the more evidence the better. However, the reality of everyday practice is far from ideal. It is still a big challenge for doctors to get the right amount and type of information at the right situation. Operational factors often frustrate the search for relevant details, which are often dispersed across multiple sources and buried deep within databases. The search results are often unstructured, redundant and/or too comprehensive to be usable. Both Smith[79] and Mickan and Askew[80] mentioned the information need from doctors' viewpoint in the information overload era: individual related patient information is needed. So it is important to consider information filters as a means to provide relevant personalised information to doctors.

2.5 Knowledge base

In this section the semantic web, ontology, OWL, the meaning of knowledge bases and databases will be introduced, and the reason why ontology-based knowledge base was chosen to organise bio-health information will be explained.

2.5.1 The semantic web and ontologies

The Semantic web was conceived by Tim Berners-Lee, the inventor of the World Wide Web, the Uniform Resource Identifier (URI), Hypertext Transfer Protocol (HTTP), and Hypertext Markup Language (HTML)[81, 82]. The major objective of the semantic web is to improve data reuse, data sharing and data integration in the data-centric new generation web[82] The World Wide Web is document-centric, where a document is the basic processing unit. In the semantic web, data is processed automatically, voluminously, precisely and efficiently. "The Semantic Web is a vision for the future of the web, in which information is given explicit meaning, making it easier for machines to automatically process and integrate information available on the web"[83]. The semantic web is a better annotated version of the existing web; i.e., data is well-defined and linked. The semantic web is evolving and is the current direction of web research, providing a universally accessible platform that allows data to be shared and processed by automated tools as well as by people, in order to reach its full potential[84]

However, semantic web technologies are still in their infancy[81], and there are not yet enough tangible higher level applications. Most of the existing efforts focus more on lower layers. In the semantic era, apart from the establishment of corresponding specifications, the first thing to do is to put data into semantic web formats[82] As ontology is a key enabling technology for the semantic web[85], ontology and knowledgebase construction will become important infrastructure components for the semantic web.

The following description of what an ontology is, is provided by John Sowa:

> "The subject of ontology is the study of the categories of things that exist or may exist in some domain. The product of such a study, called an ontology, is a catalog of the types of things that are assumed to exist in a domain of interest"[86]

Thus organising things helps one not merely to retain them, but also to find[86], reuse and share them[87].

An ontology is a form of 'catalogue', in which the basic domain knowledge units and their relationships are represented and computer-usable. Ontologies are used by people, databases, and applications that need to share domain information. They encode both knowledge in a domain and knowledge that spans domains. Knowledge is reusable through ontologies[88] Knowledge sharing and reuse provided by ontology technology play key roles in information use. An ontology can be used to specify a knowledge base's structure and its classification scheme[89].

According to the description of the semantic web from W3C, a standards development organisation, there are two major parts to the semantic web: one is the common format for data interchange; the other is language that is used to record data which corresponds to real world objects[90]. The semantic web syntax is based on resource description frameworks (RDF) to represent data, and three URIs can be combined to form a RDF statement[90, 91] RDF is a generic format which can be processed easily and automatically in data interchange[81] Web ontology language (OWL) is a vocabulary extension of RDF[92]

2.5.2 Introduction to OWL

OWL is a semantic mark up language for publishing and sharing data by ontologies on the web[92, 93], supporting advanced web search, software agents and knowledge management[90] OWL is intended to be used in information processing by machines rather than understood only by human beings[83] Compared with the extensible markup language (XML), RDF, and RDF-Schema, OWL has more semantic representation ability, and thus OWL goes beyond those languages in representing machine interpretable content[83] OWL is a vocabulary extension of RDF and is based on the DAML+OIL web ontology language, which it extends[83, 92] OWL ontologies can be encoded in different syntactic forms including RDF/XML[92, 93], which is defined by XML syntax for RDF[94]. An introduction to RDF is found in Appendix A.

There are three sublanguages of OWL: OWL Lite, OWL DL (description logic) and OWL Full (the complete OWL language). Detailed introductions to the three sublanguages can be found in Appendix A. We chose OWL DL as the ontology language in the knowledge base prototype construction, as it provides computational properties which can be used in reasoning systems[92]

OWL is used to represent vocabularies and their relationships in an ontology, and this representation is directly machine-usable[83] OWL has already been integrated as a plug-in with Protégé, which is a popular & recognised ontology editor from Stanford University. In Chapter 3 there will be a detailed introduction to this tool. There is also a platform to support ontology modelling via the Protégé-OWL editor[11]. Protégé-OWL has a user-friendly interface to define class, property, instance and their relationships following OWL syntax. The OWL ontology from the Protégé-OWL platform is encoded in XML. There is no barrier for information exchange in XML format between different systems or different computers[93] OWL statements make the ontology-based knowledge base consistent, extensible and computer-readable. Aranguren and colleagues[95] conducted a recasting of Gene Ontology by OWL and have proved these advantages.

The above provides a background introduction to the semantic web, ontology and OWL. Next I clarify what I mean by knowledge base and database before I justify my choice for this study's implementation.

2.5.3 Databases

Definition of 'database' from the High-Tech Dictionary[96]:

> 1. A large collection of data organized for rapid search and retrieval. 2. A program that manages data, and can be used to store, retrieve, and sort information.

'Database' refers to a collection of organised data in this thesis, which is the underlying meaning, rather than a strict literal definition. "A datum is a single observational point that characterizes a relationship. It generally can be regarded as the value of a specific parameter for a particular object at a given point in time"[18](p64).

2.5.4 Knowledge bases

Definition of 'knowledge base' from Lemaire[97]:

> A knowledge base is a model, a structure to store, organize and access a set of knowledge. Knowledge base is the result of a modelling process.

This definition is focused more on the function that the knowledge base can achieve. I do not intend to give a universal definition of the term 'knowledge base'; rather, and more simply, I intend to explain what I mean when I use the term in the thesis: i.e., a knowledge base is a structure to store, to organise and to retrieve computerised domain knowledge, including facts and rules.

A knowledge base 1) has inference capability by integrating a specific application domain knowledge within a database; and 2) can provide sharing, ease of maintenance and reusability of domain knowledge[98](p2). Compared with a database, a knowledge base has more capabilities in presenting, sharing, reusing and processing complex domain knowledge.

In this thesis 'knowledge base' is used as the general term, whereas 'knowledge base model' specifically refers to the logic specification for the knowledge base. 'Knowledge base prototype' is used to refer to the physical artefact of the logical specification for a knowledge base.

2.5.5 Ontology-based knowledge base

The goal of this research is to explore the feasibility of providing formal, structured and relevant bio-health information in a simulated clinical system. There are varieties of methods that could be used to provide bio-health information: for example dynamic links and tool support, which includes: databases, knowledge bases, retrieval tools etc.

Although dynamic links avoid content update, the links still need to be updated, and users need to choose the ideal links, which can be a burden for them. Another potential disadvantage related with dynamic linking is related to medico-legal reasons. For example there is a legal requirement to preserve and track what was the referenced knowledge fact at a particular time for a particular user. Retrieval tools have the same problem: searching for and then selecting correct results increases the load on the users' information use. The aim of this research is to provide the final relevant knowledge facts, providing a one-stop-shop experience, rather than a portal.

There are some requirements in providing formal, structured and relevant bio-health information for doctors' daily practice: the information should be provided in an active and intelligent way (i.e. in an automatic, context sensitive way); the content should be computer-readable, meaningful (i.e., semantic integrity), consistent and extensible. An ontology-based knowledge base is known to satisfy many of these requirements: with characteristics from both ontology (such as logic and consistent structure and explicitness) and knowledge base (representation capability, computer processability). It was therefore chosen as the knowledge management tool. We chose OWL as the ontology language because, as described in 2.5.2: it has explicit representation ability, more expressive semantics, and the representation of content is machine interpretable[83]

OWL statements make the ontology-based knowledge base consistent, extensible and computer-readable, computer-processable, and to have potential for communication and integration. It is unlikely that the knowledge base prototype we are trying to construct will work as a module of an isolated EHR system. Such a thing would be expensive and limited in value. It is more feasible that the knowledge would be communicated, and the prototype will be one of several key components in a complete system; meanwhile the knowledge base has all characteristics of OWL statements.

This is the rationale for the decisions to use an ontology-based knowledge base to provide bio-health information. There follows a brief review of how the research relates to the EHR.

2.6 Electronic Health Record (EHR)

Biological information, such as sequence variation information plays a key role in understanding disease, although the precise connection between biological information and clinical manifestations is far from clear, and hence poorly used in current clinical practice[3, 61] If biological information can be accessed easily by doctors when they are using patients' clinical information, it may broaden clinical care support and may provide potential for understanding the relationships between biological information and disease. Before anything like this can be achieved in practice, however, it is necessary to determine how this can be done, and what might be the means for delivering such benefits.

It is only relatively recently that doctors have begun to use computers directly as part of their note-taking, reading and work processes. According to Benson in the UK, "in 1975 the health centre at Ottery St. Mary, near Exeter, became the world's first paperless general practice"[99], and it was much later to develop clinical functionality rather than the more administrative tasks of repeat prescriptions and appointments. Whereas now virtually every GP uses EHR's and systems[99], it is still rare to find complete systems within secondary care[100] A new phenomenon, particularly in the US is the advent of the personal health record (PHR). There is now a pervasive trend towards using systems for supporting care. EHRs are used more by clinical staff, whereas the PHR is used by patients themselves to enter and update their personal health data. The control of the content in a PHR is therefore seen more to be the patient's responsibility.

In this research the EHR is targeted because of the clinical focus. However, there are an increasing number of technical information standards which seek to define content or/and function or communication of EHR systems, for example the HL7 Electronic Health Record System Functional Model, EN13606, CCR etc. These standards are applicable (at least in part) to both types of record system. It is through these increasingly standards-based systems that bio-health information will be delivered, accessed and used during the consultation between doctor and patient.

2.6.1 Meaning of EHR in this thesis

I used both the term 'EHR' and 'EHR system' (EHR-s) in this thesis. There are definitions of both terms in Shortliffe and Cimino's book[101](p448):

> "An electronic health record (EHR) is a repository of electronically maintained information about an individual's lifetime health status and health care, stored such that it can serve the multiple legitimate users of the record."
>
> "A computer-based patient-record (EHR) system adds information management tools to provide clinical reminders and alerts, linkages with knowledge sources for health care decision support, and analysis of aggregate data both for care management and for research."

In this thesis 'EHR' refers to all electronic bio-medical and health related data of an individual patient, excluding finance data; 'EHR-s' is a management tool or platform for capturing, storing, displaying, retrieving, transferring and managing electronic individuals' electronic health records.

The 'EHR-s prototype' is another term used in this thesis. The definition of a prototype system from Shortliffe and colleagues' book[18](p208) is:

> "Prototype systems are working models that exhibit the essential features of the system under development. Users develop a realistic idea of what the system will look like, how it will work, and what it will do."

'EHR-s prototype' is used to refer to an EHR working model, which can be demonstrated in the same way as an EHR-s; it has a subset of functions sufficient for this research but not all the functions found in a commercial product.

2.6.2 EHR-s in the clinical setting

EHR-s is still maturing and hence any review is incomplete and the results suggest much more can be done. For example 1) there is no EHR-s that can seamlessly integrate data or can coordinate processes across the entire continuous procedures of health care services; 2) there remain challenges from technology, human resources and organisational perspectives in EHR-s development and application[102] (p10); 3) the results of a systematic review of electronic medical records (EMR) to improve physician performance and patient care, are quite conservative[103].

However, the trend is for more and more of these systems to emerge. Arguments in their favour include the facts that they are much easier and more flexible to access, manage, communicate with and use for storage than paper is. Indeed, the first recommendation from the Committee on Improving the Patient Record, Institute of Medicine in the US is "health care professionals and organisations adopt the computer-based patient record as the standard for medical and all other records related to patient care"[102](p180). In the UK there is a National Programme for IT to bring modern computer systems and technology into the National Health Service (NHS), and electronic patient records are part of this programme[104, 105] National EHR architecture models have emerged in the US and Australia[106] The US is ambitious to offer most Americans EHRs before 2014[106] Although existing EHR-s are not perfect, EHR-s or EHR components in a bigger system will play a key role in clinical setting, and especially in doctors' consultation with patients. The trend means that EHRs are the most likely candidate medium for displaying/accessing/viewing biological information along with other clinical data. In the next section the major structure of an EHR, the requirements of an EHR-s and major challenges for an EHR-s will be introduced.

2.6.3 Structure of an EHR and requirements of EHR-s

The patients' health record has a long history, arguably from the earliest days of clinical practice. The EHR is a digital version of the patients' health record. Although there are some distinct characteristics for EHR particularly, paper versions of the health record and EHR have common content and characteristics. Currently with the emergence of EHR there is a trend to move from a healthcare provider-centred record to a more patient-centred record[107](p90), and from an administrative record to a clinical-use record.

There are some milestones in the history of the health record. The Problem-oriented Health Record (POMR) and Subjective, Objective, Assessment and Plan (SOAP) data categorisation were proposed by Weed in the 1960s[107](p101), while the History, Observations, Assessment, Plan (HOAP) was proposed by Donnelly in 1992[107](p101). Van de Velde & Degoulet[107] (p118) viewed the structure of EHRs in four different ways: a longitudinal life-long record; a problem-oriented record; an episode-based record and a pragmatic record.

Van de Velde & Degoulet [107](p118) summarised the gold standard for attributes of a computer-based health record system:

> "a problem list; to measure the patient's health status and functional levels; to document clinical reasoning and rationale; a longitudinal view that provides timely linkages with other patient records; guaranteed confidentiality, privacy, and audit trails; continuous access for authorized users; support for simultaneous multiple user views; direct data entry by physicians; support for practitioners in measuring or managing costs and improving quality; flexibility in supporting existing or evolving needs of clinical specialties"

Some of these attributes can be used as reference in building the EHR-s prototype.

Most of the current EHR-s are passive; the so called active EHR-s[107](p91) will increasingly interact with users (doctors or other professionals) and so offer more support for direct patient care. For this to happen some challenges need to be overcome, an important one being to connect with formalised knowledge sources to support the doctors' work[107](p93).

In the next section the available EHR tools for this research will be introduced.

2.6.4 Available EHR tools

There are many EHR tools, of which OpenEHR[108], INDIVO (formerly named PING)[109], EHR ontology[110] and HealthFrame[111] were selected as candidates for this research.

In OpenEHR there are complete and detailed specifications for the EHR[108] However, at the time my work was carried out there was no available tool from OpenEHR. Although there was a JAVA core application[112], it still lacked a usable interface. There are several other coherent tools for OpenEHR, developed by Ocean Informatics[113], such as Template Designer, Archetype Editor etc. However, none of the tools were designed to be used directly during a clinical consultation. Archetype Editor is a tool for clinical data specifications (i.e. archetypes). Template Designer is a user friendly tool with archetype-enabled data entry. Both Archetype Editor and Template Designer are important tools. However, neither of the tools could be used directly for our purpose.

INDIVO is a patient-driven open source digital health records platform[109, 114]. This is a project developed in cooperation between Harvard Medical School, MIT and Children's Hospital Boston and the Dossia Consortium. The INDIVO system is a complete and secure copy of personal medical records, across sites and over time. However, INDIVO is difficult to configure and to get the different versions of tools to cooperate with properly without conflicts.

The EHR ontology is based on OpenEHR, but the ontology is at a higher level and it is hard to use directly[110].

HealthFrame is a commercial user friendly system which is based on CCR[111] This is a tool aiming at organising entire family members' lifelong health care and wellness information. It includes all health-related information about the family member, from medical records from doctors and hospitals, to family history and personal life style. It is a personal health information hub, which is meaningful in keeping personal health information complete and correct. There is a free trial version with time limitation for evaluation. Although it is an ideal end user system, the source code was not available for our study.

Giving the limitations of time and available technical support and the cost of proprietary systems it was decided to build a system prototype based on an international standard.

2.6.5 Consideration of target user

There are many potential users of EHR systems[115], such as health care professionals and administrative staff, including specialists, general practitioners, nurses, laboratory technicians, biomedical researchers, medical students, administrative staff and patients etc. It is impossible to meet every user's demand in this study. The focus is to offer biological information in a clinical setting and to target non-specialist doctors, for example general practitioners, who might need biological information support during their consultation with patients with genetic diseases. The demands of other specialists such as biomedical researchers, can be readily satisfied by a large number of existing biological information databases. It is assumed that, compared with biological information experts, the non-specialist doctor will need more support with basic terms, such as a specific mutation names, rather than detailed biological information. This consideration will influence the content, structure and format of the knowledge base organisation.

In the next section the candidate resources and tools will be reviewed. These resources and tools are potentially useful in knowledge base prototype construction.

2.7 Biological information resources and tools

In this section: 'tool', refers to the service from an electronic source such as a database; while 'resource', refers to the content.

In this research different types of information about cystic fibrosis have been considered: scientific publications, gene sequence data, genetic information and concept descriptions. I have been pragmatic in the choice of tools.

The inclusion criteria for tools were: (1) domain coverage, which is focus; (2) free access, which is financial friendly and convenient; (3) maintained by a recognised organisation, which is credible; (4) evidence of updates, which is updated[5, 116-118].The exclusion criteria were: (1) protein data resources, which are not in the scope of this research; (2) genome resources; (3) resources without qualifiers for effective retrieval, which would not be efficient in use; (4) resources that can be processed further only by natural language processing; (5) resources for patients only; (6) commercial web sites, which could conflict with commercial bias.

I selected two books[119, 120], and the following tools: PubMed[121], Nucleotide[68], EMBL-EBI[70, 122], DDBJ[69, 123], OMIM[124, 125], Cystic Fibrosis Mutation Database (CFMDB)[126], MeSH[127], ICD-10[128], GO[56], UMLS[57], The Cochrane Library[129], and CF web pages/sites. PubMed, Nucleotide, OMIM and MeSH are all services provided by the National Centre for Biotechnology Information (NCBI) in the US. Although the interfaces are similar, the main content and focus are different. There was overlap between databases, for example between EMBL, Nucleotide and DDBJ. Some of the tools will be introduced in detail (Table 2-1), most of the others and comparison of the different gene related resources with respect to cystic fibrosis were introduced in a published paper[130]

GO and UMLS are references for knowledge base hierarchies; The Cochrane Library is used to form clinical requirement questions; other resources are used to acquire knowledge manually for knowledge base construction. ICD-10, GO and UMLS have different perspectives for description of diseases, biological concepts and medical related concepts. They are good references relating to the structure of the basic concepts in knowledge base construction.

Table 2- 1 General introduction of some tools used in this research

Name	Properties/functions
Books[119, 120]	Main body of CF knowledge; health information resources
GO[56]	Three structured, controlled vocabularies about gene products (biological processes, cellular components and molecular functions in a species-independent manner)
UMLS[57]	To facilitate the development of computer systems that behave as if they "understand" the meaning of the language of biomedicine and health
The Cochrane Library[129]	Supplies high quality and updated evidence to inform people providing and receiving care, and those responsible for research, teaching, funding and administration at all levels
CF web pages/sites	From Mayo Clinic and NLM, Human Genome Project Information, web sites from Oxford University and so on[66, 131-136]

Background facts about cystic fibrosis and why it was chosen as the exemplar condition will be introduced in the next section.

2.8 Cystic fibrosis

To carry out the research in a concrete way, cystic fibrosis was chosen as the exemplar condition.

2.8.1 Cystic fibrosis as an exemplar condition

Cystic fibrosis is a monogenetic disease. Cystic fibrosis transmembrane conductance regulator (CFTR) gene mutations cause the condition. Monogenetic disease is caused by changes or mutations that occur in the DNA sequence of one gene[66] Polygenetic diseases are due to the interactions of multiple genes and environmental factors[65]. Compared with polygenetic diseases such as heart disease, high blood pressure, Alzheimer's disease, arthritis, diabetes, cancer, and obesity, there are far fewer variables in considering the relationship between diseases and related factors in monogenetic diseases (these include cystic fibrosis, sickle cell anemia, Marfan syndrome, Huntington's disease, and hereditary hemochromatosis etc.)[66]

The incidence of cystic fibrosis in white people is relatively high. Cystic fibrosis is the most common profoundly life-shortening genetic disease in white populations[120](p5). In the US it occurs in around 1/3300 white births, 1/15300 black births, and 1/32000 Asian-American births. Cystic fibrosis is carried by approximately 3% of the white population[133] It is a lifelong disease and the life expectancy of cystic fibrosis patients has increased from an expectation of one year in 1960 to a median survival age of about 40 years in 2002[137] The longer life expectancy means that the cystic fibrosis patient will have more chances to be seen by doctors outside of specialist clinics.

2.8.2 Why choose cystic fibrosis

There are several reasons for choosing cystic fibrosis as an exemplar:

1) Cystic fibrosis involves multiple systems and associated symptoms, especially respiratory, digestive and reproductive systems. The nature of the condition will impact on a number of clinical services, involving both specialist and generalist.

2) Normally the symptoms present from an early age and persist throughout the patient's whole life. The chronic and persistent symptoms require consistent support from doctors. This will require 'continuity of care' and the associated records and record systems to support the patient over time.

3) New information, including biological information will emerge over the patient's lifetime and new information will need to be integrated into health care service and therefore will be introduced to doctors (and to patients) during the long term health care support process.

4) As cystic fibrosis has a relative high incidence and it is the most common life-threatening genetic disease in the UK, it has been studied for a long time and there is a relatively mature knowledge body. In CF many relationships between genotype and phenotype are known although it is still far from complete. Even so it means that there are more available resources for constructing a prototype knowledge base so as to test the research hypothesis.

5) Cystic fibrosis is a monogenetic disease, which is much simpler than polygenetic disease and therefore is useful for developing prototype solutions. The long term goal of this research is to integrate different types of knowledge seamlessly into the clinical setting, but we are 'walking before attempting to run'

6) Given that the totality of relationships between bio and health information are incomplete, it is useful if a platform can be constructed that will give the clinician opportunity to consider whether or not there is any new link between the data types presented. The prototype system is not a diagnosis system; rather it is a vehicle for joint presentation of different types of information.

7) Due to what is already known, the life expectancy of the cystic fibrosis patient has increased dramatically and survival expectancy is now far beyond the teenage years. Health care support for individual patients therefore will last longer than previously and new co-morbidities will appear and new links will be made. The range of services and the generalist nature of the clinicians seeing cystic fibrosis patients can be expected to increase and it is anticipated that they will need better tools than what are presently available as new problems emerge.

2.8.3 The molecular and pathological mechanisms of cystic fibrosis

Cystic fibrosis is an autosomal recessive monogenetic disorder. A defective gene- CFTR- alters a protein that regulates the normal movement of salt (sodium chloride) in and out of cells. Under normal conditions, CFTR protein plays a role as a chloride channel pumping chloride ions out of the cell. In cystic fibrosis, the chloride conductance is increased within cells; in the sweat glands, cation is prevented from absorption because the chloride cannot be absorbed as well[138, 139](p54). The fluids secreted are higher in sodium and chloride concentration, with increased salt. This leads to abnormal 'salty' secretions which promote bacterial colonisation, and result in thick and sticky secretions in the respiratory and digestive tracts, as well as in the reproductive system, and in salty sweat on the skin[134] This thick and sticky mucus clogs the lungs and leads to severe lung infections. The thick secretions also clog the pancreas, preventing digestive enzymes from reaching the intestines to help break down and absorb food[131]

Symptoms can include weight loss, troublesome coughs, repeated chest infections, salty sweat and abnormal stools[132]. Other common problems encountered in cystic fibrosis include nasal polyps, rectal prolapse, cirrhosis and diabetes mellitus, meconium ileus and congenital bilateral absence of the vas deferens in males[140]

2.8.4 Cystic fibrosis transmembrane conductance regulator (CFTR)

The CFTR gene is located in chromosome 7q31 and has 27 exons. The biological vocabulary is listed in the Appendix J for reference. CFTR protein- membrane bound glycoprotein, contains 1480 amino acids with a molecular weight of 168 kDa. The protein is comprised of two, six-span membrane-bound regions, and its primary role is to act as a chloride channel[126, 140]. There were 1571 CFTR gene mutations reported, including missense, frameshift, splicing, nonsense, deletion and sequence variation. The most common mutations are ΔF508 (70%), G542X (3%) and G551D (2%)[126, 140]. ΔF508: deletion of 3 nucleotides between 1652 and 1655 (positions of CFTR nucleotides sequence), results in loss of Phenylalanine at 508 (position of CFTR amino acids sequence)[126] The loss of amino acid leads to CFTR protein alteration, which causes the pathological changes of cystic fibrosis (2.8.3).

The genotype-phenotype relationships in cystic fibrosis are complex, and are affected by many factors: pollution, smoking, bacterial infection, malnutrition, and certain therapeutic agents. From a genetic view, within all the CFTR gene mutations, the delta F508 mutation is not only the most frequently encountered but also the most severe genetic lesion for homozygotes[141] The CFTR genotype is related to the patients' pancreatic status, about 85% of cystic fibrosis patients with pancreatic insufficiency and about 15% with pancreatic sufficiency; the correlations between the CFTR genotype and pulmonary, liver, and gastrointestinal expression are debatable[142] Although much research about the relationship between genotype and phenotype of CFTR was undertaken, their precise relationship is not well understood[137]

2.8.5 Cystic fibrosis gene therapy

Gene therapy is a technique for correcting defective genes which are causes of disease[66] The most common therapy is to insert 'normal' genes to replace 'abnormal' genes. A vector is used to deliver the 'normal' gene. There are viral vectors and non-viral vectors in gene therapy. Viral vectors used in cystic fibrosis gene therapy include adenovirus and adeno-associated virus[143]. Cationic lipids or polymers are non-viral vectors[144] Viral vectors are unsuited to repeat dosing, while non-viral approaches appear to be more suitable to repeat dosing, although they are less effective[145] Three collaborative research groups in the UK have found liposome-mediated CFTR genes can be transferred into the nasal epithelium[146]. This treatment can correct the chloride defects, but not sodium defects, with no toxic problems[146]. Gene therapy research promises some hope in future for cystic fibrosis patients and has made some progress in basic research[147]. However, so far none of the clinical investigations using adenovirus or adeno-associated virus, cationic lipids or polymers has shown a persistent correction of the ion transport defects in cystic fibrosis patients[144], and none of these trials have actually become tangible therapeutic benefit[147].

2.9 Summary

This chapter has presented the reason for undertaking this research, and a review of related work. A description of the available technology involved in this research has been given, as well as a brief introduction to the possible biological candidate tools. The exemplar condition cystic fibrosis was introduced together with the reasons for using it in this study. In this chapter, the rationale, the theoretical basis and the pragmatic constraints of the research have also been given. The following chapters will outline the research methodology.

Chapter 3 Research methodology: an overview

This chapter is a brief outline of the research methodology. It provides an overview of the 1) research question; 2) conceptual framework; 3) research strategy, including abstract and concrete solutions; 4) description of detailed solutions and the reasons for these solutions; 5) introduction of the research evaluation; 6) requirements and functions of the concrete parts of the solutions; 7) reasons for choosing CCR as the standard and 8) introduction to the development environments.

This is a proof of concept research study and the intention is to deliver biological information through a simulated clinical system. The definition for 'proof of concept' comes from the Technology Source Archives[148]: "A proof of concept is a test of an idea made by building a prototype of the application. It is an innovative, scale-down version of the system you intend to develop. In order to create a prototype, you require tools, skills, knowledge, and design specifications." The reasons for doing this research and the background have been presented in the literature review chapter. Here I restate the research question to introduce the research methodology.

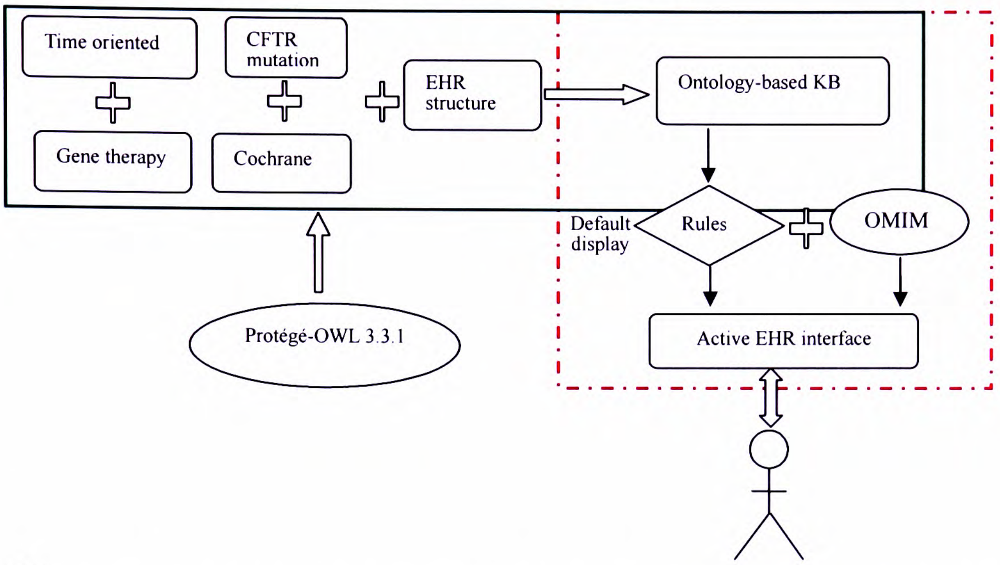
3.1 Research question

The research question is: 'to what extent is it feasible to deliver formal and structured biological information associated with a subset of health information through a standards-based clinical record'?

3.2 Conceptual framework

Oates' definition about a conceptual framework states: "A conceptual framework makes explicit how you structure your thinking about your research topic and the process undertaken"[149](p34). The following context discusses the processes used in tackling the research question. A knowledge base model was designed to organise and to represent a subset of bio-health information for cystic fibrosis. A CCR-based EHR system prototype was built to simulate the clinical record system and to make the content and structure of the knowledge base accessible. Hypothetical test records were used to show personalised bio-health information through the EHR prototype.

Figure 3-1 presents the conceptual framework, which relates to the knowledge base prototype construction and the intended connection via the record interface. This represents an overview of the method for tackling the research question. The detailed processes, i.e. the ontology construction methods and the EHR prototype and connection are dealt with in Chapter 4 and 5 respectively.



Key
Green rectangle: knowledge base prototype (KB); red rectangle: EHR system prototype; the overlap part is connection between KB and EHR.

Figure 3-1 Overview of the conceptual framework showing the physical artefacts

The general research approach is based on two literature reviews of 1) existing research in bio-health informatics and 2) the testing of a selection of electronic tools and resources. The latter review focused on selecting and comparing candidate electronic knowledge sources for bio-health information and cystic fibrosis. This was published[130] and has been included in the literature review.

3.3 Research strategy

According to Oates' definition of 'research strategy', which states "A strategy is your overall approach to answering your research question"[149](p35), this is a design and creation[149](p108) research study. "The design and creation research strategy focuses on developing new IT products, also called artefacts, including constructs, models, methods and instantiations"[149](p108). The data collection method in this research has involved using documents which include published journal papers, conference papers, textbooks, and online resources.

In this research the major work has been to create logical specifications and then to construct the designed physical artefacts according to the logical specifications. The physical artefacts construction process gives feedback to the logical specifications. Figure 3-2 shows the abstract and concrete parts of the work.

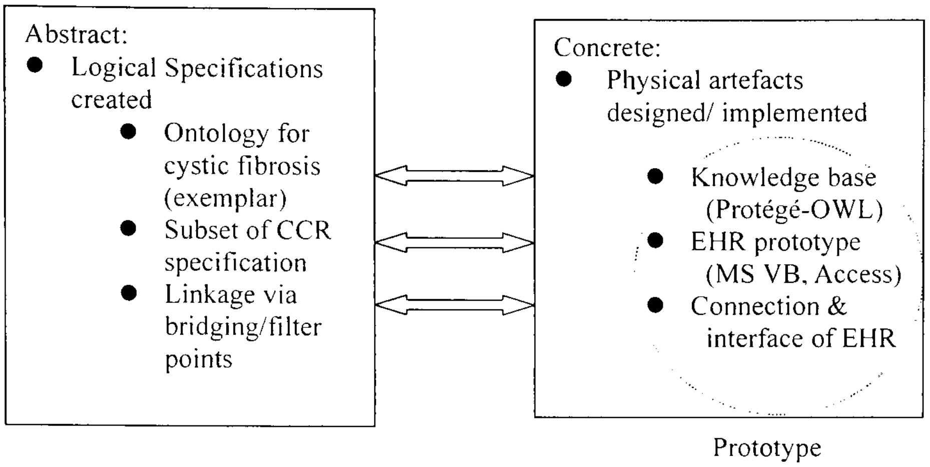

Abstract:
● Logical Specifications created
● Ontology for cystic fibrosis (exemplar)
● Subset of CCR specification
● Linkage via bridging/filter points
Concrete:
● Physical artefacts designed/ implemented
● Knowledge base (Protégé-OWL)
● EHR prototype (MS VB, Access)
● Connection & interface of EHR
Prototype

Figure 3- 2 Overview of the research strategy

(Note: specifications on the left side guide construction of physical artefacts on the right side and construction provides feedback to improve specifications)

Three major specifications have been created: 1) an ontology specification for cystic fibrosis, which is used to guide the knowledge base construction; 2) a subset of the CCR specification, which is used to build the EHR prototype; 3) linkages between biological and health information, which are used in the connection and interface of the EHR prototype.

The specifications are embodied in a complete prototype, which comprises the three physical artefacts from Figure 3-2. Meanwhile, the physical artefacts are used to show what the research is, to show what can be achieved in this research, to help communicate with peer researchers and to give feedback to improve the specifications.

3.4 Description of the research solutions

3.4.1 Detailed research solutions

The **solutions** for these objectives included[150]:

1) A **knowledge base prototype** designed to hold and to represent biological and health related information for persons having the condition, cystic fibrosis[151] The relationships which are represented are those which have been shown by domain research to exhibit significantly higher than normal level probability of manifestation. For example, an adolescent patient with cystic fibrosis has a

higher than normal probability of delayed puberty. Thus, these relationships provide 'candidate' observations concerning the type of patients not necessarily the facts about the individual patient. The major challenge here is to devise a methodology and structure for the systematic representation of bio-health research findings into an ontology based on the Protégé-OWL tool. These methods provide a paradigm which may be used for the creation of other bio-health ontologies. It needs to be noted that the requirements for integrating the bio-health information into the EHR necessitate extending what one might recognise as a conventional domain ontology by the addition of EHR structural features into the ontology.

2) To build a simulated clinical system i.e. an **Electronic Health Record (EHR)** system prototype. This is a conventional EHR system based upon the Continuity of Care Record (CCR) standards but enhanced as follows:
   - An extra sub-division of investigations has been added so as to be able to capture mutation results, e.g. sequence variations.
   - A facility is provided whereby, given the accumulated demographic and clinical information concerning the individual patient, candidate facts (e.g. signs, results, mutations) are suggested as possible investigations to be considered, i.e. the supplied information from the ontology is personalised according to the patient's age, sex, ethnicity etc.

3) To provide an interface allowing the transfer of information from the Protégé-OWL knowledge base into the EHR. This presented one of the most technically difficult aspects of the study. Protégé-OWL is very powerful in the creation of ontologies, but lacks tools for integration with other systems. The solution was to design and build a tool which took a complete download of a version of the ontology. This was analysed and used to populate a conventional database which could then be used by the EHR application. These methods are extensible, i.e. they may be applicable across other clinical and biological domains (and beyond).

3.4.2 Reasons for these solutions

Each solution step was appropriate, given the PhD constraints and available resources. The **reasons for** choosing these **solutions** included:

- Cystic fibrosis is a monogenetic disease with a relative stable knowledge body; the life expectancy of cystic fibrosis patients increased from 1 year in 1960 to a median survival age of about 40 years by 2002. It is increasingly common for patients with this condition to be treated on an ongoing basis by non-specialist doctors, such as GPs, rather than more specialist clinics. This is partly the reason GPs are considered as an example set of target users.

- Organising bio-health information as an ontology makes it sharable, reusable and extensible. These attributes offer advantages for a framework which must deal with uncertain and changing bio-health information over time.

- The EHR is increasingly used in clinical settings to manage many types of health-related information; a simulated prototype is a good choice for proof of concept. Since the nature of the biological information is new compared with demographic information and clinical information in the clinical record, there is a need to design and build an “enhanced” system based on international standards which will include biological information.

- Following a standards-based approach has the potential to provide a common underlying model for many EHR systems, irrespective of implementation. This should help future communication between systems.

- Offering customised bio-health information is important to avoid increasing the information overload faced by doctors, and also meets the demand for more personalised health care. Currently sequence variations are used as biological information examples.

3.5 Research evaluation

According to Oates, “criteria for evaluation of an IT artefact include functionality, completeness, consistency, accuracy, performance, reliability, usability, accessibility, aesthetics, entertainment, fit with organization and so on”[149](P115).

The outputs of this design and creation research are a combination of models and their instantiation. More specifically, the models are the ontological model and the EHR model; the instantiations are the knowledge base prototype, the EHR prototype and the connection between the two.

For this research, functionality and accessibility were selected to evaluate the work. Other evaluation criteria mentioned above are related to the final product, however the innovations in this research are: 1) biological information has been included; 2) formal and structured biological and health information are organised into an ontology; 3) bio-health information from the ontology can be personalised through the EHR interface. This research at the current stage aims to prove feasibility, and to demonstrate that innovations can be achieved, rather than to produce a final commercial product. The evaluation is therefore into utility evaluation, and further research should consider evaluation related to usability.

There are three major parts that need to be evaluated: logical structure of the ontology; accessibility of the knowledge base through an EHR prototype; and functionality of the connection including personalised information. The evaluation is objective-based evaluation[18] (p296) and will be of the research objectives set out beforehand. A detailed description can be found in Chapter 5 (5.7).

3.6 Requirements and functions of the physical artefacts

3.6.1 Requirements of the knowledge base

Several requirements were imposed on the bio-health knowledge base (2.2.3)[33-35]:

- To have the potential to share, reuse and expand knowledge.
- To be compatible with major existing knowledge or terminology tools.
- To have the potential to map onto future EHR systems, with respect to both basic concepts and final knowledge facts as they pertain to biological and health information.

These factors influenced the design regarding technology, content, structure and methods of construction for the knowledge base prototype.

3.6.2 Contents of the knowledge base

Whereas the trigger parameters provide an important bridging function, it is, of course necessary to have sufficient content and structure to populate the model. This section outlines the scope of the knowledge base, including the content and structures in the current version of the knowledge base model, and shows the potential characteristics intended for future versions (Table 3-1). The parameters in the following table guided the construction, and the content was based on consideration of how the knowledge base prototype would be used. The whole construction process is an interactive process, which helped to make the scope and granularity clear. The Table 3-1, therefore, is a combination of content from before and after construction. More specific details, such as reasons for choosing the content and design decisions, are presented in Chapter 4.

Table 3- 1 Overview of knowledge base prototype content and structure

	Current	Future
Biological information		
Nucleotide mutation	+	
Amino acid change	+	
Protein structure		⁘
Gene mutation location	+	
Gene/protein mutation dissection	+	
Health information		
Demographic data (age, sex, ethnicity etc.)	+	
Symptoms	+	
Diagnostic tests/diagnosis	+	
Treatment facts	+	
Prognosis		⁘
Prevention		⁘
Structure		
Basic concepts	+	
Combined concepts	+	
Final knowledge facts	+	
EHR data representation structure	+	

3.6.3 Requirements of the EHR system prototype

There are many EHR systems in place and these systems are very different to each other even within the same areas of service. For example secondary and primary care systems are often very different both in function and in terms of the degree of clinical focus. A major requirement of the research is to seek a potentially generic solution that can be applied to multiple implementations and reduce the complexity of new innovations. For this reason a standards-based solution is investigated.

The EHR system prototype should provide a series of patient-centric snapshots, which can be used to simulate a longitudinal record. The prototype should have the capability of editing, displaying and storing bio-health data and have accessed to the background knowledge[107, 152]

The knowledge base prototype plays a role as a 'plug-in' to the EHR system prototype. Individual patients' parameters trigger default displays. Parameters such as age, sex, ethnicity and specific bio-health characteristics act as a "bridge" between the EHR system prototype and the knowledge base. The displayed results can be modified by changing the patient's parameters.

There are several reasons for choosing this particular set of triggers:

1) The EHR system is the principal communication interface, therefore the data items common to such systems become the primary candidates; all the chosen triggers are classic demographic data for many if not all EHR systems.

2) Cystic fibrosis is one of the most heavily researched genetic conditions, and it became clear that these particular parameters were significant to the conditions manifestation and impact the treatment delivered.

3) Finally, some of these parameters are believed to be important to many genetic conditions, and this supports the intention of providing a generic solution.

3.6.4 Functions for the EHR system prototype

The EHR system prototype is used to simulate the clinical system the proposed target users will be non-specialist doctors, for example the general practitioner (GP). Given the increasingly widening role of primary care, and the increasing longevity of the cystic fibrosis patient, GPs are envisaged as the example target in the simulated scenario. This does not preclude the extension of the user group to other clinicians or even to patients in the future.

There are of course many different doctor-patient situations. This research has been simplified by considering only patients with an existing diagnosis of cystic fibrosis. This is not a diagnosis system. However, an individual without the condition will also have to be considered, in order to show how the interface works without a cystic fibrosis presentation. The knowledge base output is displayed as a reference or reminder through the EHR interface; the recommendation function may be considered in future.

In addition to the knowledge base prototype, the EHR system prototype will comprise additional modules and functionality. Within this research the focus is on the ideal situation in the patient-doctor consultation. If the current target functions can be achieved successfully, then more complex situations may be considered in the future.

Table 3-2 is a summary of the EHR system prototype's application environment and the major functions which have been included, along with indications for future work.

Table 3- 2 Overview of EHR system prototype- application environment and functions

	Current	Future
Target users		
Non-specialist doctor (e.g. GP)	+	
Patients/& Other clinical professionals		⁘
Patient's condition		
With CF diagnosis	+	
Without CF diagnosis	+	
EHR system prototype		
Edit patients' demographic & bio-health data	+	
Snapshot patients' records	+	
Longitudinal patients' records	+	⁘
Knowledge base display		
Reference knowledge facts	+	
Reminder	+	
Recommendation		⁘
Communication		
With patient's parameters	+	
With other reference tools		⁘

3.7 Why choose CCR?

The CCR architecture is a standard specification developed by the American Society for Testing and Materials (ASTM)[12]

> "The CCR is a core data set of the most relevant administrative, demographic, and clinical information facts about a patient's healthcare, covering one or more healthcare encounters"[12]

The CCR standard is increasingly used in the US. The major characteristic of the CCR is "continuity", which will ultimately require all of a patient's record of healthcare to be documented. This concept is essential for safe care of an individual over time and will be managed by an EHR system. However, such a record will have fine-grained individual data, whereas the facts derived from the knowledge base, OntoKBCF, are more general, covering generic attributes of the population rather than individual traits.

The increasing importance of CCR and its increasing popularity in the US (currently in

America with a very low level of physician EHR adoption rate, only 17% of physicians reported using a fully functional or basic system[153]) suggest that the CCR standard will influence many implementations in the future. Furthermore, the decision to create a prototype based on CCR is supported by the fact that it has a dataset which is in some respects close to my requirements for interfacing diversely sourced data. The design which has been utilised resembles Healthframe's, which was one of the first commercial systems to be based on the CCR. CCR was adopted since, like EN13606, it is based on a formal standard. Unlike EN13606, however, it is a simpler standard with ready-made headings to structure content.

3.8 Development environments

Protégé-OWL 3.3.1[11] has been used for the knowledge base prototype construction. This is a popular and credible knowledge management tool. Protégé-OWL 3.3.1 originated at Stanford University and has been extended at Manchester University. The developers claimed there were approximately about 100000 users currently registered, including many international users[11] The active Protégé community has contributed many plug-ins and ontologies. There are two platforms: Protégé-frame and Protégé-OWL. The Protégé-OWL platform supports OWL. The Protégé-OWL editor is used as a tool for OWL ontology construction, is freely available and is a mature and reliable tool for knowledge base construction[11]

Microsoft Visual Basic (.NET 2003) and Access 2003 were used as the development tools for the EHR system prototype and to effect the connection with the knowledge base prototype.

3.9 Summary

In this chapter the research methodology has been outlined and the overview of the research approach explained. The functional requirements of the proposed applications, the general scope and the boundaries of the knowledge base prototype and EHR prototype system, the tools and the development environment have also been introduced.

Chapter 4 Construction of the knowledge base prototype

4.1 Overview

The research aim requires the construction of the knowledge base prototype, its implementation using an ontology web language tool, and delivery to the target audience of the content of the knowledge base prototype in a context-sensitive manner. The knowledge base prototype is based on an ontology and is named OntoKBCF to reflect its basis (ontology), its purpose (knowledge base) and its application (cystic fibrosis). This chapter concentrates on the design and construction of the ontology including 1) a concise rationale as to why an ontology-based knowledge base prototype is appropriate to offer bio-health information; 2) construction approach; 3) design decisions; 4) examples and 5) discussion. It explains the meaning of 'knowledge base prototype' and 'knowledge base model' to clarify these terms when they are used.

Terminology notes:

- 'knowledge base prototype' is used when a specific reference is made to OntoKBCF, while its logical specification uses 'knowledge base model';
- 'knowledge base' is used generally to refer to a structure which stores, organises and retrieves computerised domain knowledge, including facts and rules, with ontology schema and instances;
- 'class' and 'concept' are used to refer to an OWL class in OntoKBCF;
- 'basic concept', 'combined concept', 'complex concept', 'supporting concept', 'final knowledge fact' and 'knowledge fact' are terms used specifically in OntoKBCF construction; their meaning and relationships are listed in Table 4-1 and shown in Figure 4-1;
- 'concept' is used for class in the UMLS following its original usage;
- 'term' is used for class in GO and it also follows its original use.

4.2 Introduction

The long term goal of the research is to provide different types of knowledge seamlessly in a clinical setting. The ontology-based knowledge base was chosen to provide domain knowledge facts (in this example about cystic fibrosis), in order to organise bio-health information. The knowledge base prototype was then embedded into an EHR system prototype at a later stage of the research. Semantic web technology is the major technology that has been used in the knowledge base prototype construction.

Table 4- 1 Major terms used in construction of OntoKBCF

Name	Explanation & example	Reference section
Basic concept	The atomic concept from UMLS, GO or domain knowledge, such as Gly, which is an amino acid Glycine	H1.3 in Appendix H
Combined concept	Combination of basic concepts, such as Gly542, which is Gly's location- 542 in the CFTR amino acid chain	H1.3 in Appendix H
Complex concept	A subset of combined concept explained through the EHR interface, such as Patient_CF_with_Gly542X, which is a type of CF patient with a type of mutation Gly542X	H1.4 in Appendix H
Supporting concept	Consists of basic concepts and combined concepts. Used in explaining the complex concept.	
Final knowledge fact	A domain statement, represented in OntoKBCF through combination of basic concepts or combined concepts with properties and logic relationships, such as property and description of Patient_CF_with_Gly542X	Figure H2 in Appendix H
Knowledge fact	Any above concept is a subset of this one. It includes 1) hierarchy of concepts, both basic and combined concepts; 2) property description of the concept.	

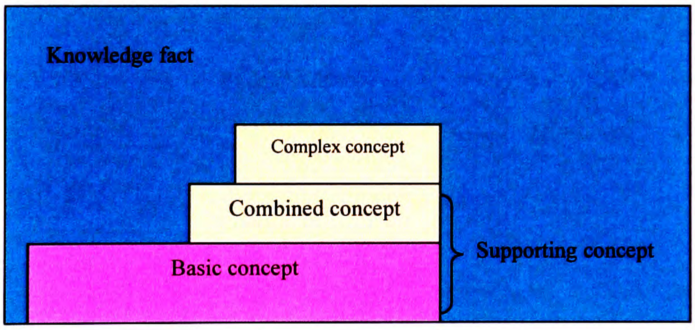


Figure 4- 1 Relationships of terms used in construction of OntoKBCF

(Note: final knowledge facts are supporting concept with description of properties)

The semantic web is the evolving direction of the current world wide web[82]. In contrast to the conventional web, which considers only machine and human processable documents, the semantic web formally considers the data within documents[82]. Semantic web technology is different from natural language processing and machine learning in artificial intelligence[82]. Semantic web technology is not to do with "real understanding" of content by machines. Rather, semantic web technology makes productive use of the "labels" associated with data elements, thus allowing machines to process the data more precisely.

There are therefore two important facets related to semantic web technology: i.e., common formats and language. Common formats are useful for data sharing, data reuse and data exchange; while language is used for representing and relating to real objects[90]. Other characteristics associated with the semantic web include: data interoperability, machine processability, and semantic understandability.

An ontology is an important infrastructure and tool in the semantic web world. It shares similar characteristics with the semantic web in that it provides the basis for a "common understanding of a domain"[85](p5). The use of an ontology provides opportunities to improve knowledge management[85](p5).

The characteristics of an ontology include:

- universal format; this is offered by a uniform resource identifier (URI);
- expression capability, which is achieved by representing data related to real objects;
- extensibility, which can be achieved by using standard resources or tools.

These three characteristics are appropriate for delivering bio-health information in a clinical setting. Consequently, the ontology-based knowledge base was selected to construct the knowledge resource.

4.3 Scope

The first task in constructing a knowledge base is to define its boundary, which makes the scope clear[154] There are several considerations in deciding the prototype's boundary. First it is necessary to consider the intended beneficiaries. Since the knowledge base prototype is intended to become a module to an EHR system prototype, the principle target users are doctors and it is therefore important to consider doctors' information needs.

4.3.1 Doctors' information needs

There are several ways to learn about the doctors' information needs including surveys[155], questionnaires[156], interviews[157], observation[158], and system log file analysis etc[159] In this study the information needs were identified from the literature for the purpose of structuring the research prototype. I used PubMed and ISI Web of Knowledge as main tools. In PubMed I combined MeSH terms (such as information service/utilization, '/' is used in PubMed to associate MeSH heading and sub-heading) and free texts (such as clinic*, information demand*, requirement*, need*, the star symbol is used as a wildcard in PubMed). The MeSH terms maybe restricted to "major MeSH topic" or to "Title" according to the retrieved records. In ISI Web of Knowledge the same free texts as in PubMed were used and the free texts were restricted to "Topic" or "Title" The retrieved papers have been reviewed in order to judge if they were relevant.

Although I was unable to find research specifically focused on doctors' information needs on cystic fibrosis, it would not be unreasonable to assume there are common needs across different specialties. According to research by Lappa[156], Smith[79] and Seol[47], the most common clinical question relates to treatment. Davies[160] conducted a systematic review of doctors' information-seeking behaviour; information need was the first theme of that paper. The paper concluded that "the top categories of information need are treatment or therapy"[160]. Arroll and colleagues[161] reported similar results, noting that questions about treatment were the most common questions asked by family physicians; with respect to content the most often raised question pertained to medical fact, followed by medical opinion and non-medical reminders[161]

Given the purpose of this research, the "'treatment' information need" and "medical facts" became the two major factors in the content selection for the knowledge base.

4.3.2 Considerations on selection of sources and tools

Apart from information needs, there are several other important factors that need to be considered in offering a knowledge resource in a clinical setting:

- available, reliable and valid knowledge sources are required;
- matching points need to be considered between the knowledge base prototype and the EHR system. Note: this includes matching points for both biological information and health information;
- compatibility with other major terminology tools at the terminology level will be

required.

4.3.3 Knowledge base prototype scope

In considering the scope, and in particular what is inside the boundary, attention is now turned to the knowledge base prototype's content. The two main content parts in the knowledge base prototype are the domain knowledge and the EHR structure. This section focuses on domain knowledge, while the EHR structure will be introduced in section 4.10.

There are four parts of the domain knowledge included in the current knowledge base. These are:

1) gene therapy with regard to cystic fibrosis[143, 146];
2) time related cystic fibrosis descriptions[119];
3) cystic fibrosis data related to the Cochrane review conclusion[129];
4) the most common CFTR mutations and their characteristics[120, 162]

Fine detail of this content is given in Appendices B to F. Three examples in section 4.9.2. 4.9.3 and 4.9.4 show how the domain content is organised within OntoKBCF For information concerning the structure and nomenclature of nucleotides, which are crucial to much of the biological information within OntoKBCF, the reader is referred to Appendix G, while the genetic vocabulary involved in OntoKBCF is given in Appendix J.

The feasibility research question required the construction of a knowledge base 'prototype', rather than a complete and comprehensive commercial product. Consequently, selected Cochrane review topics (such as airway clearance techniques for cystic fibrosis patients have short-term effects in terms of increasing mucus transport) were chosen as the starting clinical requirements, with an emphasis upon treatment. In addition, time-oriented cystic fibrosis descriptions (such as for adult male cystic fibrosis patients, manifestations include delayed puberty, sterility and bronchiectasis) were also included. Cochrane review conclusions, the most common CFTR mutations and their characteristics, and time-orientated descriptions have enabled the knowledge base prototype to be used in the prototype EHR system, and have also provided matching points with the data typically found within EHR records. In the prototype, patients' age, sex, ethnicity and clinical manifestation are taken as the matching points between the knowledge base prototype and the prototype EHR system.

As 'treatment' is the most common clinical focus and the research aim was to provide bio-health information for doctors, gene therapy became a natural join point.

With regards to factor (4), i.e. mutations, the amount of information required to be entered into the knowledge base is significantly reduced because of the intended clinical application. For although there are about 1600 CFTR mutations reported[126], approximately 70% of patient conditions can be explained by the most common CFTR mutations[120, 162] It is the mutated gene together with environmental factors that determines the manifestation of the disease. The most common CFTR mutations are included in the knowledge base prototype, in order to provide a potential foundation for future understanding of the precise mechanism between gene and disease. However, only a subset of the content in Appendices E and F has been represented in the knowledge base prototype. The criteria for choosing the mutational representation samples included are as follows:

1) For mutation among the same ethnicity group, only the higher prevalence type was selected; for example among Spanish cystic fibrosis patients, there are two types of mutation that are quite common, Gly542X and 2789+5G->A; Gly542X was represented because it has the higher prevalence.

2) Mutation types which overlapped as shown in Appendices E and F, were included.

3) The most common types were selected irrespective of other considerations, such as [Delta] Phe508. Although this type belongs to no particular ethnicity group, the prevalence of it is approximately 66%.

4) Those mutation types which were common in some ethnicity groups and were known to have clinical characteristics.

Only the uncontested knowledge facts about cystic fibrosis were represented in the knowledge base prototype. The knowledge facts may be used to answer questions about "**what**", such as:

- what is Gly542X?
- what is the hierarchy of cystic fibrosis?

Disease mechanisms, such as the way the CFTR gene mutation affects CFTR protein function, may be used to answer questions about "**how**"; reasons or explanation, whereas facts concerned with the relationship between a cystic fibrosis patient and respiratory infection, may be used to answer questions about "**why**".

Fact content about "what" was included and content about "how" and "why" were generally excluded in the current version. In part this is because some of the causal detail is contested, but also because the prototype is required only to show how bio-health information could be matched consistently, rather than being comprehensive.

Considering the requirements for the knowledge base prototype meant making pragmatic decisions concerning scope. In the light of this, the 'what' knowledge facts were considered to be sufficient to prove the concept. More complex knowledge representation has been left for future work.

4.4 Bio-health resources and their role in the ontology construction

There are three interrelated sources: basic concepts, combined concepts and final knowledge facts in OntoKBCF. The review work suggested that we construct a specific knowledge base prototype about bio-health information on cystic fibrosis and base it on existing work. At the terminology level, most of the basic concepts and their structure were selected from UMLS[57] (mostly using the Medical Subject Headings [MeSH]), their semantic types and GO[56]. Most combined concepts were derived from combining basic concepts according to the final knowledge fact. Both basic concepts and combined concepts are supporting concepts. The final knowledge facts include both health and biological content. The health information specific part was drawn mainly from The Cochrane library[129] – i.e., knowledge about cystic fibrosis treatment, while time-oriented descriptions of cystic fibrosis were from Harris and Super's book[119]. The biological part, i.e. the most common CFTR mutations, were from Orenstein[120], Zielenski[162] and the Cystic Fibrosis Mutation Database (CFMDB)[126] Other resources include several papers[143, 146, 162], web sites (such as cystic fibrosis charities[131, 132] and cystic fibrosis descriptions from Mayo Clinic service[134] and Merck Manual[133]) and

Nomenclature for the Description of Sequence Variations[163, 164]

4.4.1 Methods to obtain domain knowledge

The major methods used in obtaining domain knowledge in the OntoKBCF construction include 1) self directed learning, including on line education material, literature and cystic fibrosis societies,, 2) consultation with a human genetics expert and 3) discussion with supervisors and peer students. And the whole process (to learn, to understand, to consult, to discuss, to construct and to revise) is iterative.

The author of OntoKBCF has a Bachelor's degree of Medicine, and has been trained systematically under medical education. This is an important foundation from which to understand the health aspects of cystic fibrosis. Biochemistry and Cell Biology were two major basic medical subjects in her medical education. In preparation for the research, she learned Molecular Biology by herself from a number of resources, including text books, online training and tutorials. These three subjects are important to understand the biological aspects of cystic fibrosis.

The human genetics' expert was consulted about molecular genetics during the construction of OntoKBCF. The expert has a MSc and a PhD in Human Genetics and has worked for a Molecular Biology & Genetics laboratory in John Hopkins University for three years. Her role was to explain and to confirm any biological confusion that the author met during the construction.

4.5 Rationale for the approach

Note: Much of this section relates to the structure and nomenclature of nucleotides and amino acids. For those unclear on this subject it is suggested that Appendix G is consulted.

The general idea of the hierarchy construction has to do with external knowledge fact analysis, using a process of dissection. Reconstruction work starts from basic concepts, and then the basic concept is modified step by step until the basic concept is turned into a meaningful composite (a combined concept some combined concepts are complex concepts), representing a subject for the final knowledge fact within the prototype. For example, if a patient with Gly542X mutation needs to be represented, it takes several steps to achieve:

- '**Gly**' stands for Glycine, a type of amino acid; another abbreviation name is '**G**', however the one letter abbreviation name is not used in OntoKBCF since it can cause confusion with one of the nucleotides 'G'. This is a basic concept.
- '**X**' is the nonsense codon (cross reference Appendix J), which terminates the translation; this is a basic concept.
- ***'Gly542'*** is defined as an amino acid location in the human CFTR protein; this is a combined concept.
- ***'Gly542X'*** is defined as an amino acid substitution in human CFTR protein; this is a combined concept.
- ***'Patient_CF_with_Gly542X'*** is defined as a cystic fibrosis patient with the amino acid change. This is a combined concept and also a complex concept, which will be explained through the EHR interface (further details are provided in section 4.10.2 and Table 4-15).

Every later concept uses the former ones in its representation. The final concept may be used as the subject to describe characteristics for patients with this type of amino acid change. When the user uses the final knowledge fact, the former basic concepts can all be accessed by dissecting the more complex and specific concepts.

Figure 4-2 represents the general ideas behind analysis and construction of OntoKBCF.

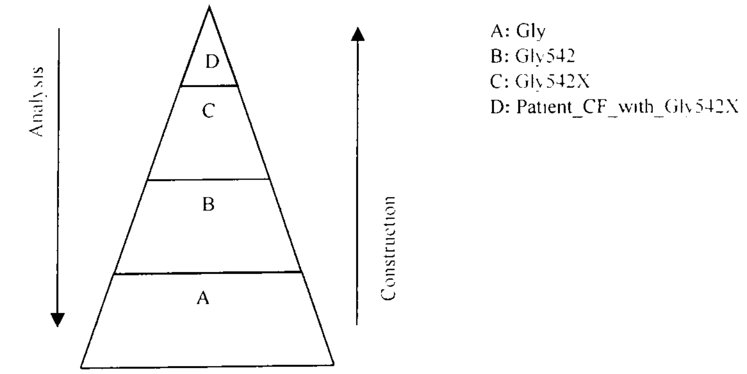


Figure 4- 2 General analysis and construction ideas for OntoKBCF

Figure 4-3 is a graphical interpretation of knowledge fact represented in OntoKBCF. In this case the subject is a female adolescent cystic fibrosis patient, which is the intersection of three bigger ellipses (Patient with CF, Adolescent, and Female). This intersection part represents all possible properties related to female adolescent cystic fibrosis patients. The knowledge facts (the representation of the subject) describing this subject in OntoKBCF are a subset of the intersection- i.e. the pattern-filled ellipse.

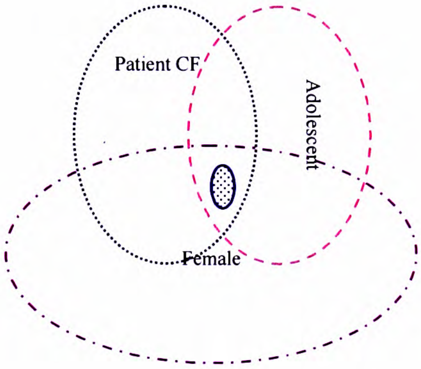


Figure 4- 3 Representation for adolescent female cystic fibrosis patient in OntoKBCF

[Note: the filled ellipse represented in OntoKBCF, compared with the whole set of properties for the subject (intersection of three bigger ellipses means all the possible properties related to an adolescent female CF patient)]

4.6 Design decisions

4.6.1 General design issues

To construct such a knowledge base would be an endless process without a realistic aim, a clear scope[154] and the practical design decisions. The former two have been introduced previously; some of the decisions taken in the construction of OntoKBCF will be introduced in this section.

Although the subject of both the knowledge base prototype and the EHR system prototype, was 'a human being', to create the potential for wider use some of the classes were specifically named, such as '***Human_CFTR_gene***' and '***Human_CFTR_gene_exon***', to differentiate them from any additional classes that may concern other animals.

There are 22 types of amino acids, so there are many more possibilities for amino acid change compared with nucleotide mutation; only the changes that fell within the scope of the work were included (detail in section 4.9.5). In contrast, as there are 8 types of nucleotides which consist of DNA and RNA, with a smaller number of possible mutations, the complete set of possible nucleotide mutations in coding DNA (cDNA) for deletion, insertion, transversion and transition were included (detail in section 4.9.5).

In the prototype the final expression of nucleotide mutation can be distinguished from amino acid change as 1) their labels contained 'minus' or 'plus', or 'Ins' or 'Del' (such as '***AA2183_minus***' or '***G621_plus_1***', '***Del394***' or '***Ins3905***'); 2) a three letter abbreviation name was used for amino acid. 'Minus' and 'plus' were used to represent nucleotide position, and examples will be given in the following paragraph. 'Ins' was used for nucleotide insertion and 'Del' is used for nucleotide deletion. In theory, not all nucleotide mutations have to be labelled with one of these 4 values – however, no such example has been found in the construction of OntoKBCF[163, 165].

4.6.2 Decisions relating to the use of Protégé-OWL

The Protégé-OWL tool provides a number of features and imposes certain restrictions which impact upon the representation of the bio-health information in the ontology. This short section describes the decisions taken with regards to the use of this tool.

Protégé-OWL 3.3.1 will not accept a class name that starts with a number, and will not accept '>' or '+' symbols as part of the class name; these are necessary symbols for nomenclature of mutation[164, 165]. '>' generally indicates a substitution of nucleotide at the DNA level (e.g. c.76A>T means the nucleotide number is 76 in the cDNA sequence and substitution is from A to T) or RNA level (e.g. r.76a>u means the nucleotide number is 76 in RNA sequence and substitution is from A to U)[163]. In OntoKBCF CFTR mutations are related only to cDNA and protein levels.

The symbols '+' or '-' indicate the nucleotide position at the beginning of the intron (+) or at the end of the intron (-). This would be represented as 'plus' and 'minus' in OntoKBCF for nucleotide substitutions. For example, in the following context I will explain the meaning of 'G621_plus_1_T': 621 is the last nucleotide for the preceding exon, G is the first nucleotide for the following intron and G is substituted by T. Therefore, in OntoKBCF this would be represented as ***G621_plus_1_T***. Amino acid substitution is handled in a similar way. '***Gly551Asp***' means that for amino acid number 551, glycine changes to aspartate.

In the prototype '***Del_A***' and '***Ins_A***' were used to represent nucleotide deletion or insertion. The name '***Ins3905_T***' means insertion in nucleotides 3905 (number) of T.

The following basic name rules are used in OntoKBCF: 1) for amino acid change (i.e. protein level), a three letter abbreviation name was used; 2) for cDNA level nucleotide change, a one letter name (such as 'A', 'T', 'C', 'G') was used; 3) a mutation name which has strictly followed the nomenclature recommendation[163, 165] has been kept in the "annotation" part for reference; 4) there were only cDNA and protein level descriptions of CFTR mutation name, not including RNA level description.

Although the nomenclature recommendation[163, 165] about mutation has not been followed strictly, it is acceptable considering the aim of the prototype which is to attempt to use the prototype to explain explicitly the meaning of mutation, instead of using a reference dictionary to map all possible mutations. The principle of the recommendation is upheld while accommodating the naming rules of the tool-Protégé-OWL 3.3.1. Another reason for not following the nomenclature recommendation strictly is related to the paper[162], book[120] and CFMDB[126] used as the knowledge resources and some mutation names follow the domain knowledge resources.

For property restrictions in the OWL representation, only someValuesFrom was used as the property restriction (appearing as 'some' in the text); it would be inappropriate to use allValuesFrom (appearing as 'only' in the text), because for 'all properties', it cannot be guaranteed that their value classes are fully specified and mutually exclusive and exhaustive. It is wiser to use "some" in most representations to leave space for future efforts to complete the description. For example: from Figure H2 (Appendix H), cystic fibrosis patient with Gly542X, the subject had the mutation property "some" Gly542X because it is possible that Gly542X is not the only mutation property for this type of patient.

According to Horridge's OWL tutorial[166], the "Domain" and "Range" conditions in the "Property" tab of the Protégé-OWL interface, are not used as constraints. They are used for class *inference*. These properties are outside of the scope of this project and consequently the "Domain" and "Range" widgets were left blank for all properties.

4.7 General construction procedures - domain knowledge structure

The general construction procedures, which were mainly influenced by Rector and colleagues' documents on Ontology Engineering[154], include:

- to look for doctors' information needs;
- to decide the knowledge base prototype's scope, main axes and granularity;
- to analyze and dissect selected knowledge facts into basic knowledge units;
- to list all involved basic knowledge units and relevant vocabularies (this step is only used if relevant concepts in UMLS or GO cannot be found) ;
- to select proper concepts/terms used in the domain according to definitions from UMLS and GO;
- to organise, create and arrange the concepts into a proper structure through an ontology editing tool, along with the original identities and alternative vocabularies;
- to represent the knowledge facts by the creation of proper representations and combination with properties by logic symbols;
- to discuss with other experts and iterate the whole construction process until the terms' positions become stable.

Figure 4-4 shows the general procedures of the knowledge base prototype construction.

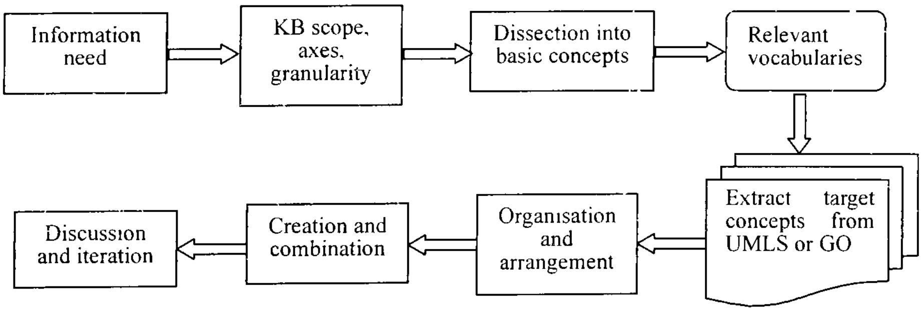

Information need
KB scope, axes, granularity
Dissection into basic concepts
Relevant vocabularies
Extract target concepts from UMLS or GO
Organisation and arrangement
Creation and combination
Discussion and iteration

Figure 4- 4 General procedures of the knowledge base prototype construction

These are the general procedures used in the prototype construction. In the next section the main axes and granularity of the knowledge base prototype and the major construction principles will be introduced.

4.8 Knowledge base prototype characteristics

4.8.1 Main axes and granularity

The main axes in the knowledge base prototype are time and problem orientation. The time-oriented axis considers how cystic fibrosis may announce itself at different ages of the subject. The problem-oriented axis here is totally different from Weeds' Problem-Oriented record[107] (p101). Here the problem-oriented axis is used to organise content to answer questions. Theses questions arise from the most common CFTR mutations, the Cochrane review conclusions and from questions about gene therapy. The entire Cochrane review conclusions about cystic fibrosis were not included. According to the Cochrane reviews' conclusion only definite and positive conclusions were considered; negative (no solid evidence to support the conclusion) and uncertainty (not sufficient evidence to support the conclusion) conclusions were excluded. The definite positive conclusions were sufficient for proving the concept at the current stage and they are clinically more robust. Negative and uncertainty conclusions can be considered for further research.

In OntoKBCF the most fine-grain level for biological information starts at nucleobase, which is the most important component for the elementary units (nucleotides) of RNA and DNA. On the other hand for health information in OntoKBCF the most fine-grain level starts from relatively atomic concepts, such as diarrhea, nausea or coughing.

It is very important to decide the scope and granularity of any knowledge base prototype. Knowledge and the splitting of knowledge facts into subunits are theoretically endless; at this stage the knowledge base prototype is only required to be constructed within the claimed scope.

Major axes are used to organise both basic concepts and combined concepts. The final knowledge facts are also organised along the main axes. The major axes integrate the knowledge base prototype as without these the knowledge base prototype would be fragmented.

4.8.2 Major construction principles

The knowledge base prototype is constructed bottom up[154], starting from the basic concepts, then combining basic concepts and finally representing knowledge facts. Dissection is used in analysing knowledge facts, and combination is used in construction. The general criterion in the dissection is based on the scope and granularity of OntoKBCF; otherwise dissection would be an endless process. Only necessary superclass concepts have been chosen rather than a detailed and complete subsets of UMLS and GO. Only the properties and constraints within the scope of the work are represented in OntoKBCF, rather than a complete and comprehensive description for each class.

Most of the hierarchies of basic concepts in the knowledge base prototype follow UMLS and GO. Relationships or classes are also added or adjusted if there are no proper choices. For example, 'sex_group' was used to connect 'population_group' and 'female'. The concept 'CFTR gene' in UMLS was adapted to 'Human CFTR gene' in OntoKBCF.

'Patient_CF' was created together with all its subclasses, i.e.

```
Patient_CF
    Patient_CF_with_age_group
    Patient_CF_with_amino_acid_change
    Patient_CF_with_neonatal_screening_for_cystic_fibrosis
    Patient_CF_with_nucleotide_mutation
    Patient_CF_with_therapy
```

4.9 Knowledge base prototype construction (OntoKBCF)

In this section, and its associated annexes, the construction of the knowledge base prototype will be explained in detail by using three pieces of knowledge facts as examples, and the organisation of biological concepts will be introduced in detail along with the major EHR structure in the prototype. Why and how EHR structure has been organised will be introduced in Chapter 5. The subjects of the cystic fibrosis domain knowledge description are cystic fibrosis patients.

4.9.1 Overview of the examples

Considering the space limitation, it is not possible to describe how every piece of knowledge fact in the knowledge base prototype has been organised. Three examples that may be seen as representative of three major parts of the knowledge base prototype content (cross reference 4.3.3 for knowledge base prototype scope) will be used. The first example is a time-oriented cystic fibrosis description; the second is about characteristics for one of the most common CFTR mutations; the third is to do with a Cochrane review conclusion. The three examples have relatively comprehensive concepts compared with other corresponding knowledge facts. I intend to give readers an outline of how to organise the domain part of knowledge facts through the three examples.

Representation font rules used in these examples (also apply to Appendices H and I)

In the following examples:

- class names are presented in **bold** (such as **Cystic_fibrosis),** and the
- classes defined locally within OntoKBCF are represented in **bold** and *italic* (such as ***Adolescent_female_CF***);
- properties are presented in *italic*, e.g. *has_diagnosis*
- restriction terms are presented underlined, e.g. some

Representation examples

Asserted Conditions (which is a *widget* on the OWL Classes tab in Protégé-OWL – 3.3.1) might appear thus:

has_diagnosis some **Cystic_fibrosis**

'some' refers to the OWL: 'someValuesFrom'.

The formal OWL representation has two dialects as shown below:

Patient_CF *has_diagnosis* someValuesFrom **Cystic_fibrosis** (W3C OWL syntax)

Patient_CF *has_diagnosis* some **Cystic_fibrosis** (Manchester OWL syntax)

The Manchester OWL syntax is used in the following description.

Please note, Example 1 will be presented next in detail; Examples 2 and 3 will only be introduced here, but the interested reader can find the full descriptions in the Appendices H and I respectively.

4.9.2 Example 1: time-oriented cystic fibrosis description

A cystic fibrosis patient will have different manifestations at different ages. Age is one of the important filters in customising bio-health information in the next stage of research. This is an example about manifestations of a female adolescent cystic fibrosis patient, and the clinical manifestations include infertility, presence of scanty cervical mucus and bronchiectasis.

4.9.2.1 Analyzing and dissecting the knowledge fact into basic concepts

Basic concepts for the first example:

- sex: female;
- age group: adolescent;
- diseases: cystic fibrosis, infertility, respiratory tract diseases-bronchiectasis;
- human being group: patient;
- quantitative concept: scanty;
- body substance: cervical mucus

4.9.2.2 Selecting concepts from UMLS

The corresponding concepts from UMLS were chosen according to the basic concepts required from 4.9.2.1 and their superclasses and UMLS's definitions. The Tables 4-2 to 4-5 organise the necessary concepts for example 1 and show their hierarchies from UMLS. The hierarchies in the tables are not exhaustive.

In UMLS there was no specific concept for patients in different disease categories. There was only one concept named **Patient** in UMLS.

Table 4- 2 Population group related concepts from UMLS

Conceptual_entity
Group
Population_group
Female
Age_group
Adolescent_age_group

Table 4- 3 Disease related concepts from UMLS

Phenomenon_or_process
Pathologic_function
Disease_or_syndrome
Diseases
Cystic_fibrosis
Infertility
Respiratory_tract_diseases
Bronchial_diseases
Bronchiectasis

Table 4- 4 Quantitative related concepts from UMLS

Conceptual_entity
Idea_or_concept
Quantitative_concept
Scanty

Table 4- 5 Body substance related concepts from UMLS

Physical_object
Substance
Body_substance
Mucous_body_substance
Cervix_mucus

4.9.2.3 Creating and organising concepts

When concepts from UMLS were organised in OntoKBCF the original identities and alternative vocabularies were kept as **comment** properties. In this case the original identities have a concept unique identifier (CUI), which can be used to track the concept over time[167] and may also be used as a "primary key" for further applications. Alternative vocabularies are synonyms. Both CUI and synonyms were recorded in the "Annotations" widget in the "OWL Classes" tab.

The new class '***Patient_CF***' was created as a subclass of '**Human_being**' to represent patients diagnosed with cystic fibrosis. A patient with cystic fibrosis has different properties, such as different age groups, different mutations and different therapies, and these were the subjects for representation of the chosen knowledge facts.

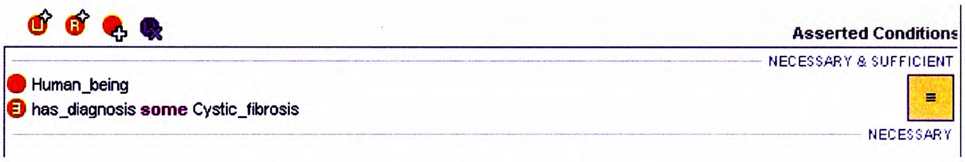
The concept '***Patient_CF***' was split further. The hierarchy for '***Patient_CF***' is shown in Table 4-6. There are several other sibling classes of '***Adolescent_CF***', such as: '***Adult_CF***', '***Child_CF***', '***Infant_CF***', and '***Infant_newborn_CF***'.

The '***Patient_CF***' was defined as the intersection of (1) **Human_being**; (2) *has_diagnosis* <u>some</u> **Cystic_fibrosis**. Figure 4-5 is the screen shot.

'*has_diagnosis*' is a property used to describe a human being with a particular disease or syndrome.

Table 4- 6 Hierarchy of *Adolescent_female_CF* in OntoKBCF

Physical_object
Human_being
Patient_CF
Patient_CF_with_age_group
Adolescent_CF
Adolescent_female_CF



Asserted Conditions

NECESSARY & SUFFICIENT

● Human_being

⊖ has_diagnosis **some** Cystic_fibrosis

NECESSARY

Figure 4- 5 Representation of ***Patient_CF***

According to the exemplar knowledge fact, female adolescent cystic fibrosis patients need to be represented. Subclasses of '***Patient_CF***' were defined as follows:

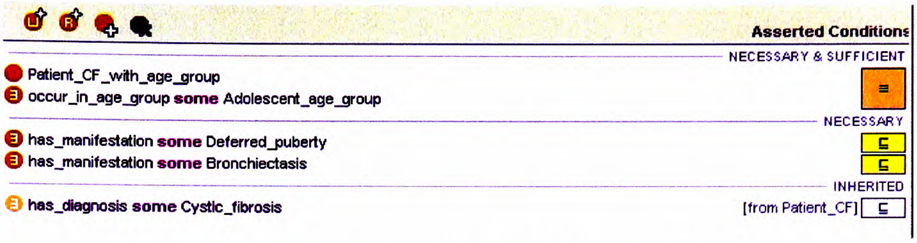
'Patient_CF_with_age_group': this class was only an abstract and interim class which was used for organising subclasses.

'Adolescent_CF' was an intersection of (1) ***Patient_CF_with_age_group***; (2) *occur_in_age_group* <u>some</u> **Adolescent_age_group;** the hierarchy is shown in Figure 4-6.

'*occur_in_age_group*' was a sub property of '*occur*'. It was a property to describe human beings in different age groups.

'Adolescent_female_CF' was an intersection of (1) ***Adolescent_CF***; (2) *occur_in_sex_group* <u>some</u> **Female** (Figure 4-7).

'occur_in_sex_group' was another sub property of '*occur*'. It was a property to describe a human being in a male or female group.



Asserted Conditions

NECESSARY & SUFFICIENT

● Patient_CF_with_age_group
occur_in_age_group **some** Adolescent_age_group ≡

NECESSARY

has_manifestation **some** Deferred_puberty ⊑
has_manifestation **some** Bronchiectasis ⊑

INHERITED

has_diagnosis **some** Cystic_fibrosis [from Patient_CF] ⊑

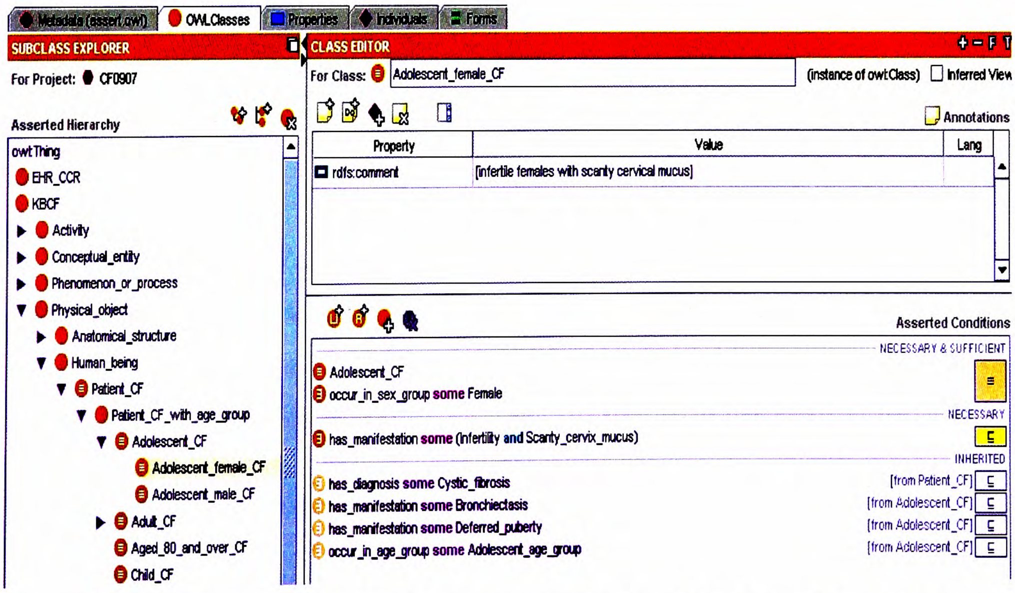
Figure 4- 6 Representation for the ***Adolescent_CF***

4.9.2.4 Representing the knowledge fact

After organising and rearranging the basic asserted hierarchy, it is now necessary to represent the final knowledge fact by combining the classes and properties. The final representation of this fact was under the '***Adolescent_female_CF***' class as a necessary condition: *has_manifestation* some *(***Infertility** and ***Scanty_cervix_mucus)***. There were also inherited descriptions from higher-level classes. Figure 4-7 shows both the representation and the hierarchical view of the ***Adolescent_female_CF***.

'has_manifestation' was a property to describe the manifestation of a human being's diseases or syndromes.

'Scanty_cervix_mucus' was defined as in intersection of (1) **Cervix_mucus**; (2) *has_quantitative_property* some **Scanty** (Figure 4-8).



For Project: ● CF0907

For Class: Adolescent_female_CF (instance of owl:Class) □ Inferred View

Property	Value	Lang
rdfs:comment	[infertile females with scanty cervical mucus]	

owl:Thing
- EHR_CCR
- KBCF
- ▶ Activity
- ▶ Conceptual_entity
- ▶ Phenomenon_or_process
- ▼ Physical_object
  - ▶ Anatomical_structure
  - ▼ Human_being
    - ▼ Patient_CF
      - ▼ Patient_CF_with_age_group
        - ▼ Adolescent_CF
          - Adolescent_female_CF
          - Adolescent_male_CF
        - ▶ Adult_CF
        - Aged_80_and_over_CF
        - Child_CF

NECESSARY & SUFFICIENT

Adolescent_CF
occur_in_sex_group **some** Female ≡

NECESSARY

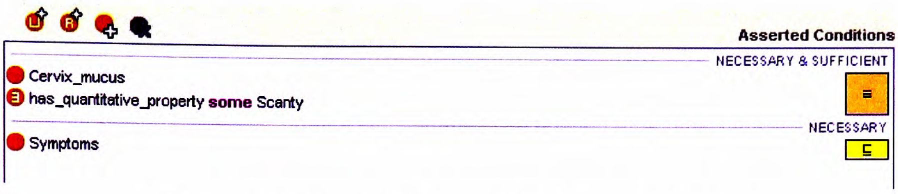
has_manifestation **some** (Infertility **and** Scanty_cervix_mucus) ⊑

INHERITED

has_diagnosis **some** Cystic_fibrosis [from Patient_CF] ⊑
has_manifestation **some** Bronchiectasis [from Adolescent_CF] ⊑
has_manifestation **some** Deferred_puberty [from Adolescent_CF] ⊑
occur_in_age_group **some** Adolescent_age_group [from Adolescent_CF] ⊑

Figure 4- 7 Whole screenshot for the representation of ***Adolescent_female_CF***

'has_quantitative_property' is a property used to describe objects with quantitative properties such as incomplete or scanty.



● Cervix_mucus
∃ has_quantitative_property **some** Scanty
● Symptoms

Figure 4- 8 Representation of ***Scanty_cervix_mucus***

4.9.3 Example 2: one of the most common CFTR mutations- Gly542X description

This example is presented in full in Appendix H. This is a description of Gly542X, one of the most common CFTR gene mutations. Gly542X is located in exon 11 of the CFTR gene and has a higher prevalence among Spanish cystic fibrosis patients. Patients with Gly542X have pancreatic insufficiency and many of them have meconium ileus.

CFTR mutations are related to ethnicity, which is another important filter used in individualising patient information. In this case Gly542X is related to Spanish patients.

The major focus illustrated in this example relates to cystic fibrosis patients with a Gly542X and includes an explanation of the mutation and relationships to ethnic characteristics and clinical manifestations.

4.9.4 Example 3: A Cochrane review topic description

This example is presented in full in Appendix I. The knowledge fact of Example 3 is: intravenous pamidronate treatment increases bone mineral density at axial sites in cystic fibrosis patients, although it can cause severe bone pain in participants not receiving corticosteroids.

Some Cochrane review conclusions about cystic fibrosis have been included in the knowledge base prototype. This is the part which can provide matching points with health information (such as symptoms). This health information has also been used in individualising patient information.

4.9.5 Overview of the biological hierarchy in OntoKBCF

In this section the structure of biological contents in OntoKBCF will be introduced. Since the major content of OntoKBCF is bio-health information on cystic fibrosis, for the health part (i.e. phenotype), most of the basic concepts and their hierarchies follow UMLS; for the biological part (i.e. genotype), most of the basic biological concepts are organised hierarchically according to common biological knowledge as found in the literature. There is no tool in the biological field corresponding to UMLS.

In this section I will introduce how the biological contents are organised in OntoKBCF, and then will introduce:

1) mutation phenomena through '***Nucleotide_mutation***', '**Amino_acid_change**' and
2) the basic components of nucleotide and protein through '**Nucleic_acid_nucleoside_or_nucleotide**' and '**Amino_acid_peptide_or_protein**'.

4.9.5.1 Major organisation of biological concepts

The major biological concepts in OntoKBCF were organised into three classes:

1) *spatial concepts* location of amino acids, nucleotides, arm of chromosome, gene exon and intron;
2) *phenomenon or process*- nucleotide mutation and amino acids translation change;
3) *physical objects*- amino acids, nucleotides and patients with different CFTR gene mutations.

The major biological facts in OntoKBCF were the location of the most common CFTR gene mutations, protein alterations, nucleotide mutations and amino acids changes, corresponding patients' characteristics (mapping points) and necessary basic concepts used to represent the mutation. The granularity started from amino acids, nucleotides and nucleobases.

4.9.5.2 Hierarchy of '*Nucleotide_mutation*' and '**Amino_acid_change**'

In this section gene mutation and amino acids change related concepts and hierarchies in OntoKBCF will be introduced. The major hierarchies of '***Nucleotide_mutation***' and '**Amino_Acid_change**' are shown in Table 4-7.

Under ‘**Nucleotide_deletion**’ all the possible subclasses in cDNA level: Del_ A, C, G, T, were listed (Table 4-8). ‘***Del394_TT***’ was classified as a subclass of ‘***Del_T***’ and its representation is shown in Figure 4-9.

‘***Nucleotide_insertion***’ was quite similar to ‘**Nucleotide_deletion**’, with Ins_A, C, G, T, u as subclasses.

Table 4- 7 Hierarchy of mutation related classes in OntoKBCF

Phenomenon_or_process
Nucleotide_mutation
Nucleotide_deletion
Nucleotide_insertion
Nucleotide_transition
Nucleotide_tranversion
Amino_acid_change
Amino_acid_deletion
Amino_acid_insertion
Amino_acid_substitution

There are four possible nucleotide transitions, ‘***A_transition_G***’, ‘***C_transition_T***’, ‘***G_transition_A***’, and ‘***T_transition_C***’. Every specific transition could be classified into one of the subclasses (Table 4-9). Figure 4-10 is an example representation of specific transition- ‘***G1717_minus_1_A***’.

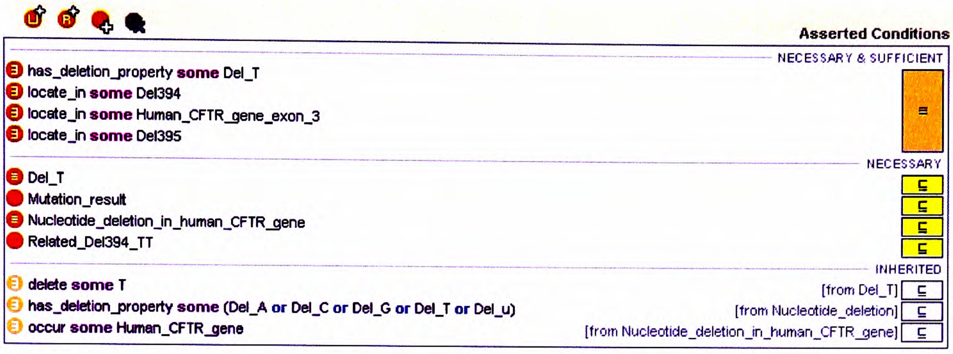
Table 4- 8 Hierarchy of **Nucleotide_deletion** and its subclasses in OntoKBCF

Phenomenon_or_process
Nucleotide_muation
Nucleotide_deletion
Del_A
AA2183_minus_G
Del_C
Del_G
Del_T
Del394_TT
Del_u
Nucleotide_deletion_in_human_CFTR_gene
AA2183_minus_G
Del394_TT

There are eight possible transversions as subclasses of ‘***Nucleotide_transversion***’. Table

4-10 shows the hierarchy detail and Figure 4-11 shows the representation of a specific tranversion-'*G621_plus_1_T'*.

The complete hierarchy for '**Amino_acid_change**' and its subclasses are shown in Table 4-11.

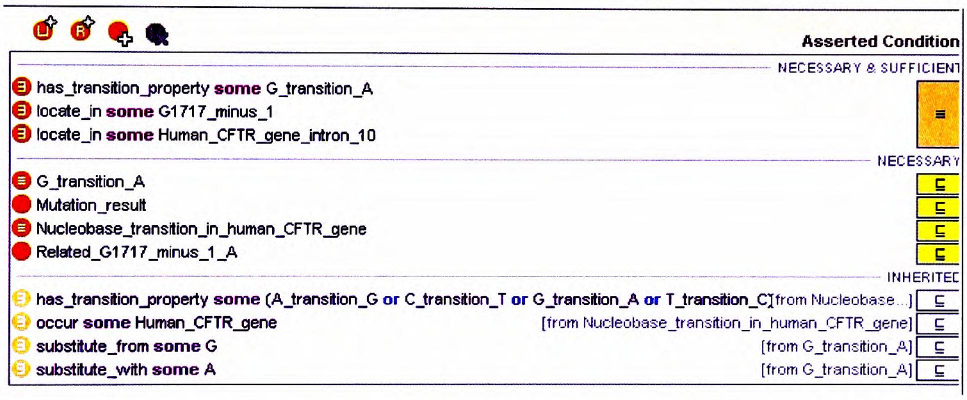


has_deletion_property **some** Del_T
locate_in **some** Del394
locate_in **some** Human_CFTR_gene_exon_3
locate_in **some** Del395

Del_T ⊑
Mutation_result ⊑
Nucleotide_deletion_in_human_CFTR_gene ⊑
Related_Del394_TT ⊑

delete **some** T [from Del_T] ⊑
has_deletion_property **some** (Del_A **or** Del_C **or** Del_G **or** Del_T **or** Del_u) [from Nucleotide_deletion] ⊑
occur **some** Human_CFTR_gene [from Nucleotide_deletion_in_human_CFTR_gene] ⊑

Figure 4- 9 Representation of ***Del394_TT*** in OntoKBCF



has_transition_property **some** G_transition_A
locate_in **some** G1717_minus_1
locate_in **some** Human_CFTR_gene_intron_10

G_transition_A ⊑
Mutation_result ⊑
Nucleobase_transition_in_human_CFTR_gene ⊑
Related_G1717_minus_1_A ⊑

has_transition_property **some** (A_transition_G **or** C_transition_T **or** G_transition_A **or** T_transition_C)[from Nucleobase...] ⊑
occur **some** Human_CFTR_gene [from Nucleobase_transition_in_human_CFTR_gene] ⊑
substitute_from **some** G [from G_transition_A] ⊑
substitute_with **some** A [from G_transition_A] ⊑

Figure 4- 10 Representation of ***G1717_minus_1_A*** in OntoKBCF

Table 4- 9 Hierarchy of ***Nucleotide_transition*** and its subclasses in OntoKBCF

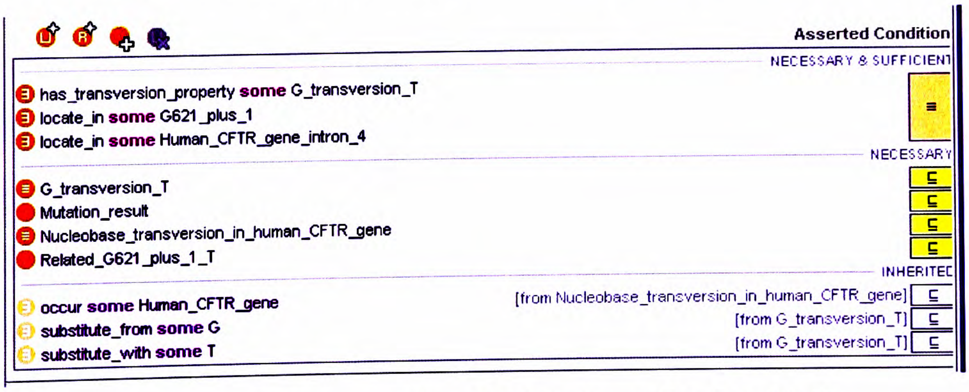
Phenomenon_or_process
Nucleotide_muation
Nucleotide_transition
A_transition_G
AA2183_minus_G
C_transition_T
G_transition_A
G1717_minus_1_A
Nucleotide_transition_in_human_CFTR_gene
AA2183_minus_G
G1717_minus_1_A
T_transition_C

Table 4- 10 Hierarchy of ***Nucleotide_transversion*** and its subclasses in OntoKBCF

Phenomenon_or_process
Nucleotide_muation
Nucleotide_transversion
General_transversion
A_transversion_C
A_transversion_T
C_transversion_A
C_transversion_G
G_transversion_C
G_transversion_T
G621_plus_1_T
T_transversion_A
T_transversion_G
Nucleotide_transversion_in_human_CFTR_gene
G621_plus_1_T

4.9.5.3 Hierarchy of '**Amino_acid_peptide_or_protein**' and '**Nucleic_acid_nucleoside_or_nucleotide**'

Basic components for amino acid and nucleotide will be introduced in this section. The general hierarchies of '**Amino_acid_peptide_or_protein**' and '**Nucleic_acid_nucleoside_or_nucleotide**' are shown in Table 4-12. '**Nonsense_codon**' and '***X***' were equivalent classes to represent translation termination codon (Table H-9). Under '**Amino_acids**' all the amino acids' three letter abbreviation names were listed.



has_transversion_property **some** G_transversion_T
locate_in **some** G621_plus_1
locate_in **some** Human_CFTR_gene_intron_4

G_transversion_T ⊑
Mutation_result ⊑
Nucleobase_transversion_in_human_CFTR_gene ⊑
Related_G621_plus_1_T ⊑

occur **some** Human_CFTR_gene [from Nucleobase_transversion_in_human_CFTR_gene] ⊑
substitute_from **some** G [from G_transversion_T] ⊑
substitute_with **some** T [from G_transversion_T] ⊑

Figure 4- 11 Representation of ***G621_plus_1_T*** in OntoKBCF

Table 4- 11 Hierarchy of **Amino_acid_change** and its subclasses in OntoKBCF

Phenomenon_or_process
Amino_acid_change
Amino_acid_deletion
Amino_acid_deletion_in_human_CFTR_protein
Delta_Phe508
Amino_acid_insertion
Amino_acid_substitution
Amino_acid_substitution_in_human_CFTR_protein
Arg553X
Asn1303Lys
Gly542X
Gly551Asp
Trp1282X

'***Nucleotide***' was used as a subclass of '**Nucleic_acid_nucleoside_or_nucleotide**' and superclass of all the nucleotides consisting of DNA (A, C, G, T) and RNA (a, c, g, u). '***Nucleobase***' was chosen to organise all the nucleotide bases with full names, such as Adenine and Cytosine. Both nucleotides and nucleobases were also classified into '**Purines_and_derivatives**' (including A, G, a, g, Adenine, Guanine) and '**Pyrimidines_and_derivatives**' (including C, T, c, u, Cytosine, Thymine and Uracil).

The detail hierarchy of 'DNA' is shown in Table 4-13.

Table 4- 12 Hierarchy of amino acid and nucleic acid in OntoKBCF

Physical_object
Substance
Chemical
Amino_acid_peptide_or_protein
Amino_acids
Human_CFTR_protein
Nucleic_acid_nucleoside_or_nucleotide
DNA
mRNA
Nonsense_codon
Nucleobase
Nucleotide
Purines_and_derivatives
Pyrimidines_and_derivatives
RNA

The preceding sections in this chapter have focused on the structure of biological related concepts. In the next section the EHR structure will be introduced.

Table 4- 13 Hierarchy of **DNA** and its subclasses in OntoKBCF

Physical_object
Substance
Chemical
Nucleic_acid_nucleoside_or_nucleotide
DNA
DNA_nonsense_codon
Exon
Human_CFTR_gene_exon
Human_CFTR_gene_exon_3
Human_CFTR_gene_exon_10
Intron
Human_CFTR_gene_intron
Human_CFTR_gene_intron_4
Human_CFTR_gene_intron_10

4.10 Structures used to facilitate the integration of OntoKBCF and an EHR

4.10.1 The nature of the requirements

OntoKBCF covered cystic fibrosis related bio-health knowledge facts. However, making available the information to the target users in a useful and context-sensitive manner within an EHR system provides a major challenge. The nature of the challenge is three-fold.

1) Within an EHR, information is generally categorised and organised under certain headings, but such categorisations are generally absent in an ontological representation of data. For example, an ontology entry for 'scanty cervical mucus' needs to be recognised as a 'symptom' in the EHR.

2) When using the Protégé-OWL it is possible, if one is experienced in its use, to identify relationships or inferences. These are of the type: for a newborn cystic fibrosis patient, the manifestation could be prolonged icterus. Such inferences may be complex, involving Boolean operators such as 'and' or 'or'.

3) Certain entries within the ontology are only understandable to those with specialist training. This applies particularly to the representation of mutations.

Generally, these are provided with explanations, which are held as relationships or properties within the ontology.

All these types of relationships or properties can be made available in the ontology, but there are no current tools which facilitate their export into an external system such as an EHR. This presented a major problem. The Protégé-OWL and its representation are not suitable for the general clinician, yet the knowledge base representation is powerful and rigorous. A bridge is required.

4.10.2 The adopted solution

OntoKBCF was expanded so as to include a set of EHR-related meta-data which could be exported into an EHR. In effect the following structure was adopted (Figure 4-12). The explanation of the related concepts will be in the following context.

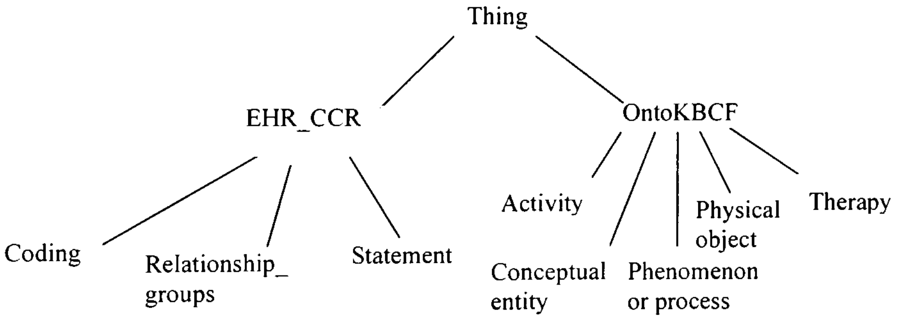


Figure 4- 12 Major structure in OntoKBCF

OntoKBCF retains its conventional structure and the EHR-CCR 'points to' entries in the ontology which have the additional relationships. CCR is used in the EHR system prototype development and it will be introduced in Chapter 5.

The major structural elements of the EHR part in OntokBCF are shown in Table 4-14. Most of subclasses of '**Coding**' (the categorisation elements) come from CCR, except for '***Mutation_result***' which is not currently a CCR category. In Table 4-14 only the major structure is listed, i.e. not all classes are fully expanded.

In construction of the ontology it was necessary to identify which leaf-node should be provided with the additional information that they may be categorised as a *Condition, Therapeutic_procedure, Mutation_result* etc. For example, '***AA2183_minus_G***', '***Delta_Phe508***', and '***Gly542X***' were all made subclasses of '***Mutation_result***'. All subclasses of '**Coding**', such as a particular problem, procedure or result, were used to represent the structure and content of the EHR system prototype.

Table 4- 14 Major structure of the EHR part in OntoKBCF

EHR_CCR
Data_representation
Coding
Problems
Conditions
Symptoms
Procedures
Diagnostic_procedure
Therapeutic_procedure
Results
Diagnostic_result
Mutation_result
Therapeutic_result
Relationship_groups
Statement

In addition to the categorisation of certain ontology class elements, '***Relationship_groups***' and '***Statement***' were used to index complex concepts, such as '***Patient_CF_with_Gly542X***' and supporting concepts in EHR system prototype.

Table 4-15 and 4-16 list some subclasses for '***Relationship_groups***' and '***Statement***' separately. The subclasses of '***Relationship_groups***' were different groups, in which the piece of complex concept and its supporting concepts (as explanation units) were grouped together under one superclass. For example under '***Related_Gly542X***' there are three subclasses, '***Patient_CF_with_Gly542X***', '***Gly542X***' and '***Gly542***'. All the three classes were defined rather than drawn directly from UMLS; the latter two were explanation units (also described as supporting concepts) for the first class, which was a complex concept and the subject with many properties and constraints for the final knowledge fact.

Only the appearance and structure of classes are introduced in this section. How these two classes would be used will be introduced in Chapter 5. Through this representation the user can get properties and constraints used to describe the final knowledge fact and can also split and trace down the meaning of subunits used for the final knowledge fact.

'***Statement***' was the collection of the subjects for final knowledge facts which could be accessed through the EHR system prototype. This is the major EHR structure in OntoKBCF.

Table 4- 15 Example groups of the ***Relationship_groups*** in OntoKBCF

EHR_CCR
Data_representation
Relationship_groups
Related_Gly542X
Gly542
Gly542X
Patient_CF_with_Gly542X
Related_AA2183_minus_G
AA2183_minus
AA2183_minus_G
Patient_CF_with_AA2183_minus_G

Table 4- 16 Example subclasses of the ***Statement*** in OntoKBCF

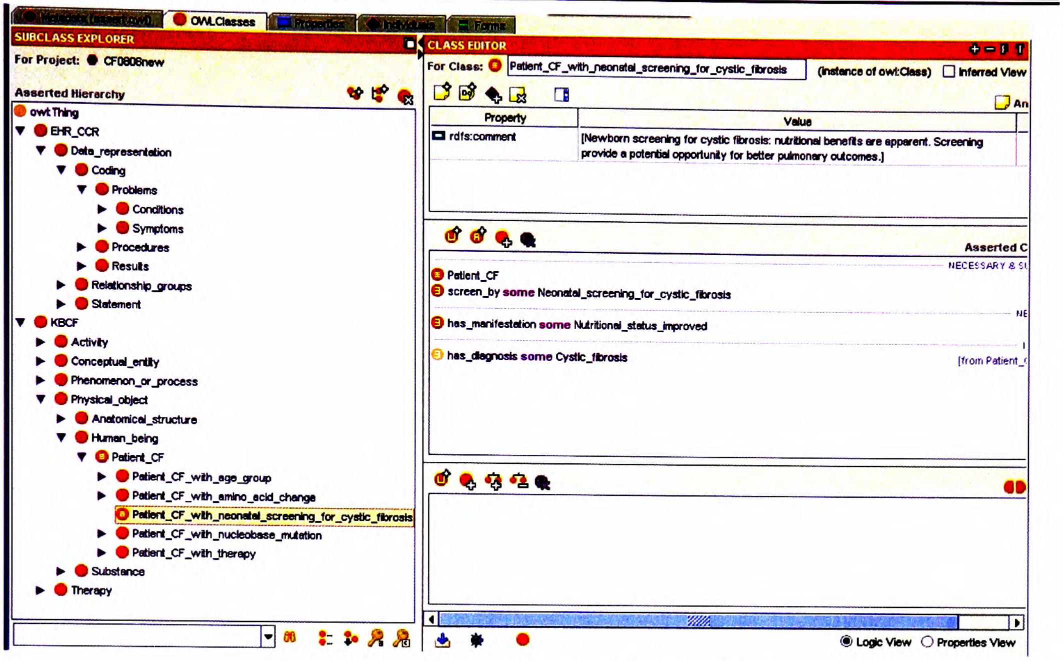
EHR_CCR
Data_representation
Statement
Patient_CF_with_AA2183_minus_G
Patient_CF_with_Arg553X
Patient_CF_with_Gly542X
Patient_CF_with_Ins3905_T
Insufficient_digestive_enzymes

4.11 Results and summary of OntoKBCF

There were two major content sections in OntoKBCF: one was for the cystic fibrosis domain and the other was for EHR structure. In the cystic fibrosis domain part concepts were organised into five classes: (1) activity; (2) conceptual entity; (3) phenomenon or process; (4) physical object; (5) therapy. In the EHR structure part the major classes were (1) coding part; (2) relationship group and (3) statement.

Figure 4-13 shows the general hierarchy of OntoKBCF and Figure 4-14 shows the general tree structure of OntoKBCF.

Table 4-17 is statistics about OntoKBCF provided by Protégé-OWL 3.3.1.



For Project: ● CF0806new

- owl:Thing
- ▼ EHR_CCR
  - ▼ Data_representation
    - ▼ Coding
      - ▼ Problems
        - ▶ Conditions
        - ▶ Symptoms
      - ▶ Procedures
      - ▶ Results
    - ▶ Relationship_groups
    - ▶ Statement
- ▼ KBCF
  - ▶ Activity
  - ▶ Conceptual_entity
  - ▶ Phenomenon_or_process
  - ▼ Physical_object
    - ▶ Anatomical_structure
    - ▼ Human_being
      - ▼ Patient_CF
        - ▶ Patient_CF_with_age_group
        - ▶ Patient_CF_with_amino_acid_change
        - Patient_CF_with_neonatal_screening_for_cystic_fibrosis
        - ▶ Patient_CF_with_nucleobase_mutation
        - ▶ Patient_CF_with_therapy
    - ▶ Substance
  - ▶ Therapy

For Class: Patient_CF_with_neonatal_screening_for_cystic_fibrosis (instance of owl:Class) □ Inferred View

Property	Value
rdfs:comment	[Newborn screening for cystic fibrosis: nutritional benefits are apparent. Screening provide a potential opportunity for better pulmonary outcomes.]

Patient_CF
screen_by some Neonatal_screening_for_cystic_fibrosis
has_manifestation some Nutritional_status_improved
has_diagnosis some Cystic_fibrosis

Figure 4- 13 General hierarchy in OntoKBCF

Table 4- 17 Statistic results about the content in OntoKBCF

Name	Number
Named classes	405
Defined named classes	112
Anonymous classes	208
Mean named parents	3
Max named parents	6
Mean siblings	9
Max siblings	20
Properties	35

The cystic fibrosis domain knowledge is organised by time and by problems. The content of OntoKBCF includes basic biological and health concepts, combined concepts and final knowledge facts. Knowledge facts include Cochrane conclusions, time-oriented description of cystic fibrosis, and the most common CFTR gene mutations description and gene therapy of cystic fibrosis.

Figure 4-15 shows what is included in OntoKBCF. In this figure the whole bio-health knowledge body is represented as a grid. However, the content represented in OntoKBCF is a small subset of the whole cystic fibrosis bio-health knowledge body.

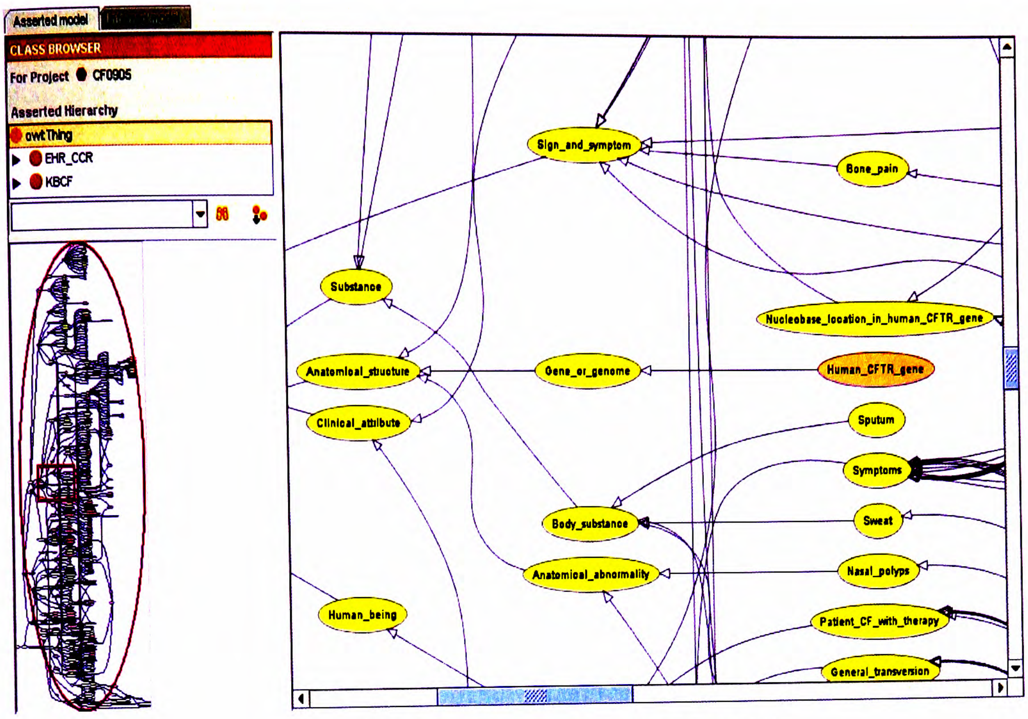


Figure 4- 14 The tree structure of OntoKBCF

(Note: the structure inside the oval on the left bottom is the whole structure of OntoKBCF and the right side part is the zoom out result of the square inside the oval)

	P	D
Bio	Mutation	Gene therapy
Health	Age/sex	Treatment/ Cochrane

Figure 4- 15 Contents represented in OntoKBCF

(Note: P = patient focused information; D = doctor focused information; Bio = biological information; Health = health information)

The user can start from biological or health knowledge fact (the most complex representation in OntoKBCF, such as knowledge fact in example 1 or 2), then trace down to the most basic biological or health concepts (such as '**Bronchiectasis**' or '**Glycine**' in Table 4-3 and Table H-1). The basic concepts and some combined biological or health concepts (such as: '***Gly542X***' in Table H-6, '***Adolescent_CF***' in Table 4-6) are both components for the final knowledge facts representation (Figure 4-6 and Figure H-2).

4.12 Discussion

It should be noted that OntoKBCF contains only a subset of cystic fibrosis bio-health knowledge facts and thus may not meet the needs of applications that operate outside the scope of the work. Limitations in the construction, mapping possibilities with upper ontologies and preparation for further steps will be discussed in the next section.

4.12.1 Upper ontology

There have been a number of attempts to develop agreed upper level ontologies, such as Medical Ontology, SnapBFO (snapshot ontologies, indexed by times) and SpanBFO (videoscopic ontology)[168]. These were not used in the construction of OntoKBCF. OntoKBCF has been constructed as a vehicle for transferring formal and structured bio-health information to a clinical setting. Thus the focus has been on pragmatic development rather than on a more philosophical formal representation. OntoKBCF is a purposive knowledge base prototype rather than a complete biomedical vocabulary. In addition, the lack of clear and complete definitions for the classes of existing upper ontologies makes it difficult at present to develop OntoKBCF according to such a framework, or to map it post hoc.

4.12.2 Problem concepts

An attempt was made to represent all the contents selected for the knowledge base prototype. However, several relatively imprecise contents could not be represented. For example, "nasal polyps, especially if recurrent" is highly relevant when diagnosing older children with cystic fibrosis. But it is difficult to represent concepts like "especially" in OWL. It is also difficult to represent the difference between relative expressions such as "very common" and "sometimes" – such expressions are quite common in clinical descriptions. Other problematic concepts include prevalence and percentages. In these difficult cases the original expression (real meaning) of every knowledge fact retained as an annotation serves to preserve their exact meaning.

4.12.3 Preparation for further steps

Both clinical and biological data will be available within the EHR prototype. In order to deliver both biological information and health information, 'mapping points' were introduced as a potential bridge between OntoKBCF and the EHR prototype. In OntoKBCF age, sex, ethnicity, mutation type and treatment are all considered to be 'mapping points' that determine how content from OntoKBCF might be displayed through the EHR prototype. Detailed content will be introduced in Chapter 5.

4.13 Summary

In this chapter, the construction of the knowledge base prototype OntoKBCF has been considered in detail. OntoKBCF includes two major parts: cystic fibrosis bio-health domain information and a structure to support or interact with the EHR. Concerning cystic fibrosis, the knowledge facts range from nucleotide mutation, amino acid change, basic biological concepts to biological and health knowledge facts of cystic fibrosis. Although horizontally the prototype is far from complete, vertically, it is able to bridge knowledge of different types, and at different levels. The description of sex, age, ethnicity, mutation and clinical manifestations provides mapping points with the EHR system prototype. Semantic web technologies offer consistent, extensible, machine processable and sharable knowledge sources. OntoKBCF is an attempt to focus upon information about the sequence variations (both nucleotides and amino acids) associated with health information about cystic fibrosis. As this is important for a complete understanding of the disease, the goal is to deliver the information and to make it usable in a clinical setting.

While recognising that the current state of the art does not permit a seamless transition from biological to clinical knowledge, the organisation of OntoKBCF is designed to provide appropriate access to biological knowledge in the clinical setting.

Chapter 5 Connection of a knowledge base prototype to a standardised electronic health record prototype

Part 1 of this research combines formal and structured biological and health information into a knowledge base for cystic fibrosis[151]. This is described in Chapter 4. Part 2 (described here) explores how this information can be made accessible via a standards-based Electronic Health Record system prototype.

Note: A review concerning EHR system (EHR-s) and the distinction between EHR and EHR-s is presented in section 2.6.

This section includes details concerning:

- Why an EHR system prototype is created, rather than re-using an existing product.
- An overview of the extensions to the existing standards, so as to include biological information.
- The processes require to integrate the ontology with a standards-based EHR application.

5.1 EHR system prototype overview

EHR-s in their various guises are popular tools in the clinical setting, as shown in section 2.6.2. These EHR-s are proprietary, and do not lend themselves to the extensions which were required; i.e. no system to the author's knowledge has been designed for the explicit purpose that we need. These additional requirements include the capability to integrate formal and structured biological and health information into EHR-s.

All computer-based systems and therefore all EHR-s are designed with specific user types in mind. For the purposes of this project the target users were non-specialist doctors, for example general practitioners. While it is true that similar techniques and general design principles would apply to other user types, the non-specialist has some specific requirements. The key element is that the specialist information (such as genetic information) could be linked to familiar information (i.e. health) by presenting the information together and providing correct interpretation of the specialist information. Furthermore, it is important to limit the amount of information to be managed by the user. This means that the EHR-s should present information in a context-sensitive manner, i.e. that appropriate information is displayed according to a patient's individual parameters. It should be noted that there has been little research thus far to bring biological information to the EHR (see section 2.6.2).

For this research, a limited number of resources have been brought together to create a knowledge base which could interface with a standard EHR-s. Most EHR-s, however, have evolved from a reducing number of suppliers (e.g. in UK general practice in 2009 there are about 10 main systems[169, 170], whereas in 1990 there were over 100[171]). It is only recently that there has been a move towards standards-based systems, and the underlying design principles of most commercially available systems are not based on formal standards or specifications but are ad hoc and unique to a specific development. These systems are not open source and their architectures are proprietary and unobtainable for interfacing with research software.

The new CEN and ISO standards such as ISO EN13606 are expected to be the basis for the record systems in the next 15-25 years, considering the standards development period[9, 10]. These standards and others are in the public domain now and are candidates for this research. Three systems were considered, e.g., OpenEHR[108], INDIVO[109], and Healthframe[111] and all of them have been reviewed in section 2.6.4. In the following material the major barriers to using them as the basis for the prototype will be introduced.

The OpenEHR utilises the CEN work, which is relatively mature, given that the original work began on the record architecture in 1992[9, 172]. There have been several developments and enhancements; the most recent undergoing balloting in ISO is in 2009. The standards work has benefited from the involvement of the OpenEHR foundation and the development of tooling from Ocean Informatics[9, 113]. However, at the time of this research there was not an available interface which could be used directly in OpenEHR.

The INDIVO system, formerly PING, developed separately from the standards organisations and also differed from many other initiatives as it was focused upon personal health records (PHR), designed to be created and operated by individual patients rather than institutions. INDIVO is open-source and is freely downloadable. However, implementation of INDIVO involves a wide range of third party tools which require considerable time if these resources are to interoperate without conflicts.

Furthermore, the initial focus of this work is intended for non-specialist doctors rather than patients, and INDIVO is not likely to set a standard which other systems would follow.

Healthframe is a commercial PHR system. Unlike INDIVO, it is standards-based and developed on CCR. However, its proprietary nature meant that the source code and architecture of Healthframe are not easily adapted. Like INDIVO, Healthframe is a PHR system to be used by patients rather than by doctors.

If we had used an existing system and attempted to extend the functionality, then we would have been constrained by the system's programming language and its architecture, both of which were potential barriers. Considering the research objective and pragmatics we were not concerned with utilising the full functionality of such systems. Therefore, it was decided to use a subset, albeit based on a standard- CCR, to prove the concept, as this was a more realistic and time saving option. The reason for choosing CCR can be cross referenced to section 3.7.

5.2 Approach of connecting the EHR system prototype to OntoKBCF

To achieve the purpose of interfacing the bio-health information and to make it accessible via an EHR the following steps were used to create the connection between the EHR and the knowledge source- OntoKBCF. Figure 5-1 shows the general procedures used in this connection implementation. Basically there are five major parts:

1) To develop the EHR system prototype [5.3].

2) To connect the EHR system prototype with OntoKBCF [5.4].

3) To restructure the patient's record from published case reports [5.5].

4) To map the knowledge in OntoKBCF with CCR labels [5.6].

5) To test the connection [5.7].

5.3 Development of the EHR system prototype

5.3.1 Selection of labels from CCR

It was never the aim of this project to build a fully functioning GP EHR-s. As this research is to show proof of concept it is not necessary to include every aspect of CCR, and therefore I chose only those parts required in the EHR system prototype to achieve the specified aim. The analysis of the CCR was concerned with only the clinical and demographic parts, but not the administrative ones (management or financial information), although all three are covered in CCR[12].

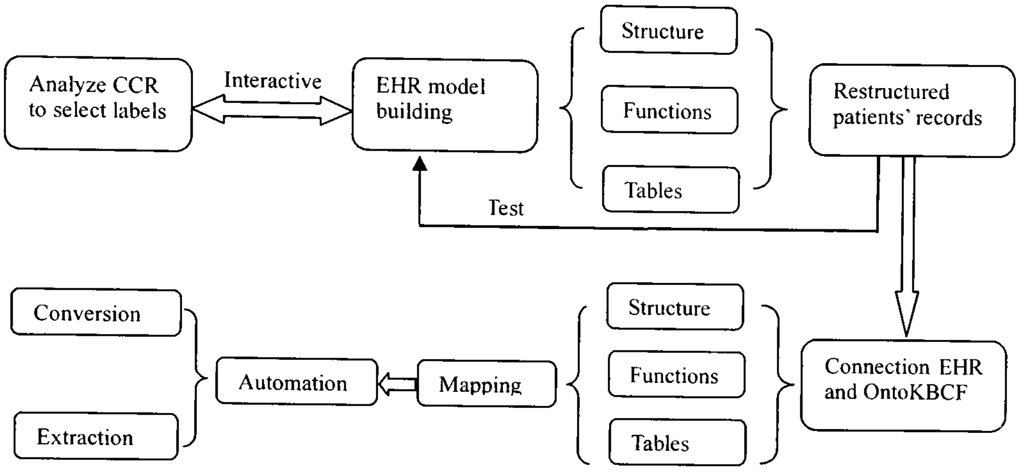


Figure 5- 1 Flow chart for EHR/OntoKBCF connection prototype construction

CCR uses labels (or CCR Attributes). As the aim was to connect OntoKBCF with the EHR system prototype, i.e., to make knowledge facts in OntoKBCF actively accessible through the EHR interface, it was critical to think of this application whilst building the EHR system prototype. Therefore the basic principle was to select labels which were applicable to the knowledge within OntoKBCF and to those found in patients' records.

There are three components in the CCR specification: the *CCR header*, the *CCR body* and the *CCR footer* respectively. The clinical labels are from the CCR *body*; the demographic data labels are from the CCR *footer*; patient's ID and consulting date/time are from the *header*.

Clinical labels are selected to organise the bio-health information, and demographic data is organised under demographic labels. Note: In this context *clinical* means medical related rather than administrative and demographic information. In this chapter this covers both medical and biological information. The specific labels will be introduced in detail in the next section.

5.3.2 Structure and functions of the EHR system prototype

Generally there are three major parts in the EHR system prototype:

- Ontological data entry (to load different versions of OntoKBCF)
- The patients' entry (to choose the patient from a list with some basic demographic data)
- EHR system prototype interface (to display the selected patient's record)

The data contained in the system may be classified as demographic data and bio-health data.

5.3.3 Demographic data

Demographic data includes: 'Surname', 'Given name', 'Date of birth', 'Sex', 'Ethnicity' and 'Age group'.

Certain demographic data items have particular prominence since they interact with the data held in the knowledge base and are required for the presentation of possible conditions/tests in a context-sensitive manner. These items are:

- Sex. Certain conditions are only relative to a particular sex.
- Age. Certain conditions are relative to a particular age group. This may also relate to sex, e.g. adolescent female cystic fibrosis patient.
- Ethnicity. Certain mutation types may be relatively highly prevalent among certain ethnic groups.

It is recognised that the latter two have certain flaws in that 'age groups' are not standardised. For example, the term *baby* may have different ranges in different contexts and *ethnicity* is even more contentious. However, previous research findings did utilise these terms and the precise standardisation and specification of these categories followed the definitions in UMLS.

5.3.4 Bio-health data

Bio-health data includes 'Problems', 'Procedures' and 'Results'.

'Problems' include 'Conditions' and 'Symptoms'.

'Procedures' include 'Diagnostic Procedures' and 'Therapeutic Procedures'.

'Results' include 'Diagnostic Results', 'Mutation Results' and 'Therapeutic Results'.

Except for 'Mutation Results', all other labels followed the CCR's definition and term selection. 'Mutation Results' was added to record and present the patient's specific sequence variations. 'Mutation Results' is a result of a laboratory test.

Figure 5-2 shows the major structure and functions of the EHR system prototype and also shows information flow in the EHR system prototype. The structure and function of the OntoKBCF related part, i.e. the connection part, will be introduced next.

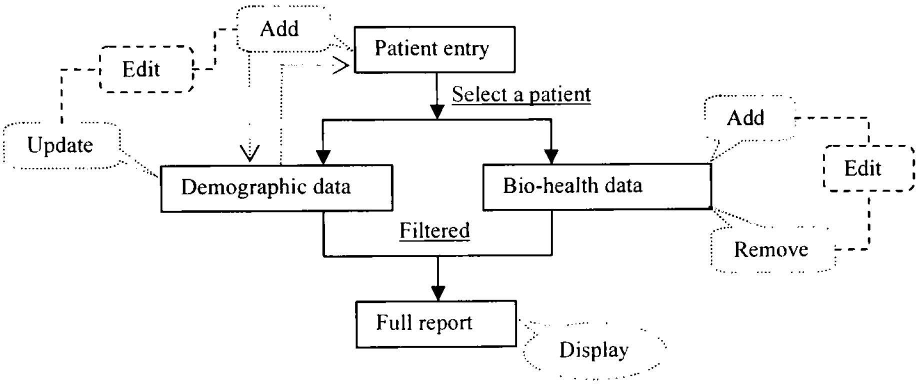


Figure 5- 2 EHR system prototype interface-major structure and functions

(Note: solid line and rectangles are for structure; dashed line, arrows and callouts are for functions)

5.3.5 Candidate facts from OntoKBCF

This section displays case-sensitive candidate knowledge facts from OntoKBCF. All the contents from OntoKBCF had been organised into five specific subparts: 'Age/Sex related', 'Ethnic related', 'Mutation related', 'Diagnostic procedure related' and 'Therapy related'.

The contents in the five specific subparts were originally represented in OntoKBCF. Only the content that is related to the patient's demographic data but not listed in the bio-health data would be displayed in these five boxes (the screenshots are in section 5.7- Verification). This is because the current filters include patient's age, sex, ethnicity and specific bio-health data, for example symptom, mutation, condition etc. Each item in the boxes could be triggered to show the hierarchies in the tree view parts and if the item belonged to one of the bio-health data categories the item could be added to the patient's bio-health data part (after the user confirms it). 'Delta_F508', for example, (a type of mutation) is an item displayed in the 'ethnic related' box. The user can decide if 'Delta_F508' should be added to the 'mutation result'.

5.3.6 Tree view

This is the part to display the hierarchies from OntoKBCF and the natural language interpretation of the representation. There are three subparts:

- The 'Original statement' is the original representation of knowledge fact from OntoKBCF;
- 'Reference hierarchy for key concepts' displayed hierarchies of the key concepts (i.e. selected items from one of the OntoKBCF boxes, described in section 5.3.5) in the original statement from OntoKBCF and their relationships;
- 'Interpretation of the original statement' was a natural language interpretation of the original statement in OntoKBCF.

Another important characteristic in 'Reference hierarchy for key concepts' is the hierarchies, which include not only the item of interest, but may also include some related concepts (i.e. supporting concepts) which are used in explaining that item. Currently only the location information for mutation concepts has been included as related concepts. The hierarchical view and the interpretation box are different methods to support users in understanding/interpreting the knowledge facts from OntoKBCF. The architecture of the EHR system prototype is summarised in Figure 5-3. The screen shots shown in section 5.7 give a detailed view of the EHR system prototype interface. Appendix M is a table with some main coding design decisions in building the EHR prototype for reference.

5.4 Connecting the EHR system prototype to OntoKBCF

5.4.1 Context of the connection

The major focus of this research is to construct an ontology for cystic fibrosis containing formal, structured and associated bio-health information and to make this information accessible in a context sensitive manner to doctors through a standards-based EHR application. It was expected that the interface between the ontology and the EHR would involve the creation of queries within the EHR application and that such queries would be sent to the ontology. The returning 'information' would then be interpreted and presented to the user (Figure 5-4).

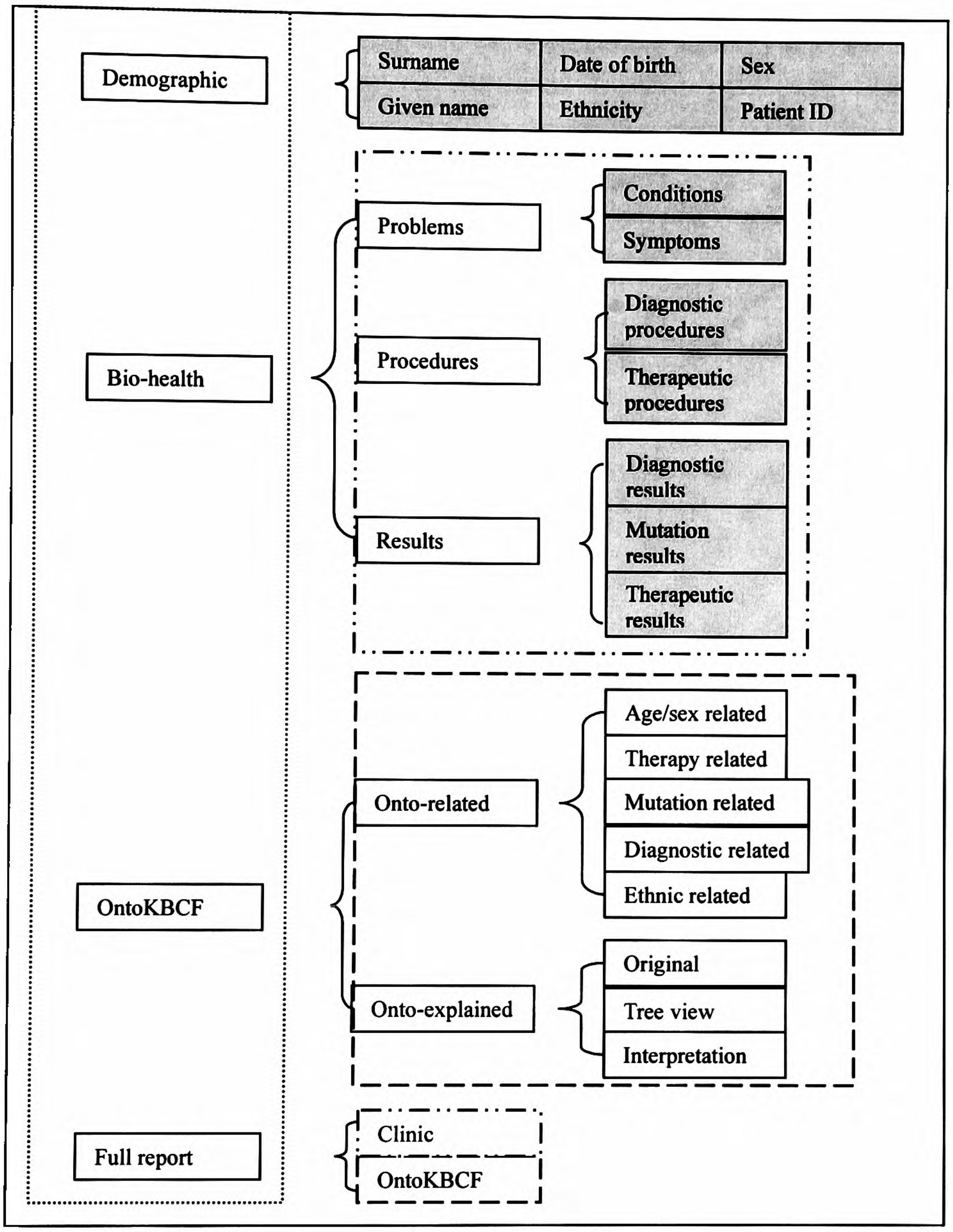


Figure 5- 3 EHR system prototype interface architecture

(The parts could be edited are in grey background; most bio-health labels and demographic labels from CCR; 'Clinic' under 'Full report' includes both demographic data and bio-health data.)

Unfortunately, when the available tools were investigated it was found that none were able to deal effectively with the types of queries that were necessary. When we tried different ways to query OntoKBCF or try to get the knowledge facts from OntoKBCF, we faced many different types of barriers. The barriers to the queries will be discussed in detail in Chapter 6. This lack of effective querying tools was brought to the attention and confirmed by Protégé-OWL experts, the developers at Stanford University and Manchester University. The only available option was to effectively download the content of the ontology for rendering into more accessible representations.

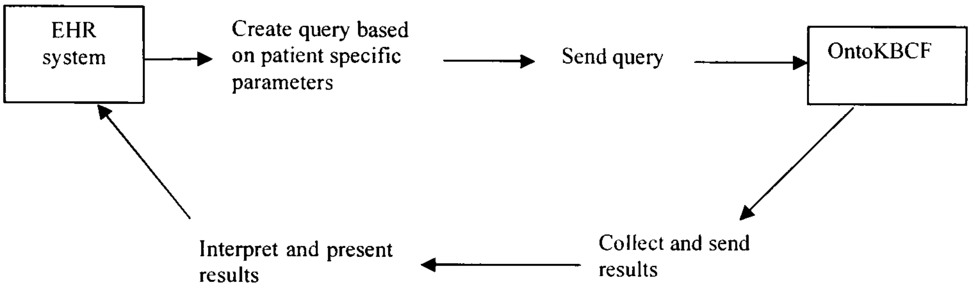

EHR system
Create query based on patient specific parameters
Send query
OntoKBCF
Collect and send results
Interpret and present results

Figure 5- 4 Ideal information flow between the EHR system and OntoKBCF

5.4.2 Possible options for the connection

The decision to move to a download configuration meant that at any given time we needed to take an image or 'snapshot' of the ontological representation of the knowledge. Downloading of the ontology content can be carried out in two ways:

- A function in the Protégé-OWL tool may be used to extract content as an XML representation. It is understood that this is the preferred option of researchers at the Mayo Clinic in the US.
- A function in the protégé-OWL tool is to download the 'raw' data from the ontology as a database format. This is the method adopted by this project since additional material which is placed in the ontology is allowed to be manupilated.

Whichever of the two methods is adopted there are serious losses of information. The two possible paths to information loss are described in Figure 5-5. These are not the only processes involved in information loss; all the possible processes related with information loss in this project will be discussed in Chapter 6.

5.4.3 Our solution

The raw data in the Figure 5-5 is converted from Protégé-OWL to Microsoft's (MS)

Access format; the knowledge facts in OntoKBCF are mapped to the CCR labels; and then data in the Access format is extracted into the bio-health tables and the OntoKBCF related tables used in the database for the EHR system prototype. The extracted knowledge facts from OntoKBCF are classes. The knowledge facts are related to specific characteristics, such as age, sex, ethnicity, diagnostic procedures, CFTR mutations and therapeutic procedures. The process has been done by a combination of manual and automatic means. In the manual process the content is analyzed in OntoKBCF directly and only the automatic process involves the converted table directly.

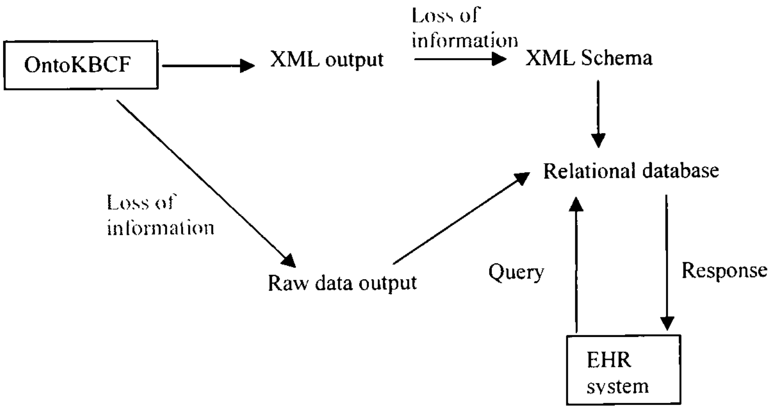


Figure 5-5 Information loss between the EHR system and OntoKBCF by the two solutions

The solution included:

1) To design what functionalities could be achieved in the connection, according to the contents in OntoKBCF, and the data included in the EHR system prototype interface.

2) To design the OntoKBCF related parts in the EHR system prototype interface.

3) Then to select the relevant knowledge facts from OntoKBCF according to the desired functionalities and to organise them into separate tables.

Figure 5-6 presents the general view of the connection. The major tables of the EHR/OntoKBCF connection prototype could be classified into three groups: patient related, bio-health related and OntoKBCF related tables. The data source for the first group is from the published case reports and details will be described in section 5.5; the data sources of the latter two are mapping results between OntoKBCF and the CCR labels, which will be described in section 5.6. This part of the work was conducted manually first and automatically later.

There are two major procedures in raw data processing: to convert the data from Protégé-OWL format to MS Access table; to extract corresponding bio-health tables and OntoKBCF related tables from the converted Access table.

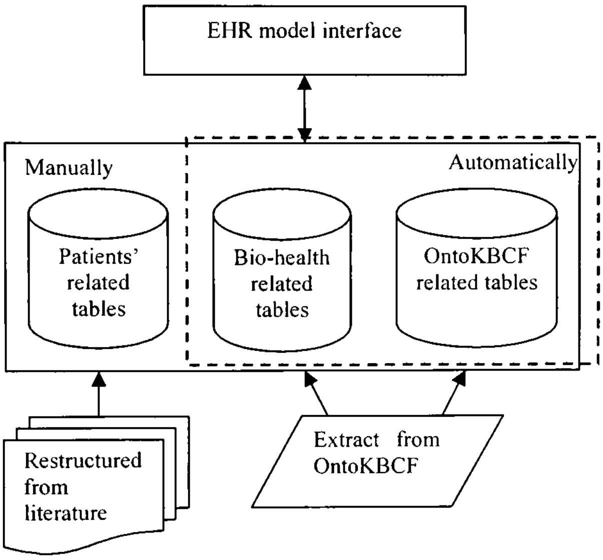


Figure 5-6 General view of the EHR/OntoKBCF connection prototype

5.4.4 Conversion table- from Protégé-OWL to MS Access table

Based on help documents from the Protégé community[11], the parameters were set to convert OntoKBCF from current format into a relational database format (a Microsoft Access table). The different bio-health and OntoKBCF related tables would be extracted from this converted table in the automation process.

There are several major characteristics in the data fields in this converted table (details in Appendix L). All these characteristics are critical clues for further automatic extracting. All the Asserted Hierarchies (i.e. 'classes' names), Annotations, Object properties, and Annotations for Property, were displayed in the original format in the table. However, the Asserted Conditions were displayed as an internal string in the form: "@_:A0", or "@_:A632" etc. Note: Asserted Hierarchies, Annotations, Object properties, Annotations for Property, Asserted Conditions are all panel names on the Protégé-OWL interface.

5.4.5 Extraction from the converted Access table

The aim of this part of the work was to extract corresponding tables from OntoKBCF automatically, according to different requirements. The first step in the programming was to analyse the manually built tables and to decide which ones could be extracted automatically. Nothing could be done from OntoKBCF directly with respect to the patient-related tables; however, all other tables could be extracted partly or totally from the original converted table.

In addition to the connection of the two types of data, OntoKBCF could also be used as a data source for the EHR system prototype's bio-health data: all the candidate bio-health data lists are from OntoKBCF. The procedure for using OntoKBCF as the data sources for the EHR system prototype, and the data sources are mapping results between the two components, which will be introduced in section 5.6.

5.4.6 Automation of extraction

The manual method is a way to prove the concept; however, it is far from ideal, requiring much laborious work without the assistance of automation. The extraction part can be automated, and in Appendix N we list the major programming decisions involved with automation of the extraction. However, at this time, not all the results achieved by the manual method can be done automatically. Results from manual and automatic processes will be compared in section 5.8.

5.5 Test patient records

To test the functions of this OntoKBCF/EHR connection prototype it was necessary to load some test records. These records were anonymous. Given that the patient had cystic fibrosis, using the interface to the ontology, the possible indications and suggestions for further investigation could be displayed. This section will introduce how the test records were restructured.

5.5.1 Test record source

A major source of test records came from case reports in PubMed. Twenty four human cystic fibrosis genetics case reports with free full text were retrieved until Jan. 21st 2008. Case reports with mutation types which overlapped with the content in OntoKBCF were selected. If there were several case reports with the same mutation type, then the one with the most detailed clinical description was selected.

Finally three published case reports[173-175] were selected, one of which included four separate cases. I chose three out of four from the same publication because the last case was almost the same as one of the previous selected three. In total, five test records were reconstructed using this method.

5.5.2 Test record restructuring

The above resources presented data in a textual manner, and in each case it was necessary to deconstruct the information into a number of factors. These factors were contained in the ontology, for example: age, sex, ethnicity, CFTR mutation type and other clinical description, i.e. symptom, diagnostic procedure and medication etc.

The method used was to restructure the test records from published case reports first, then to add several hypothetic parameters such as name, or date of birth. Normally in the published case reports patients' age, sex and ethnic origin would be indicated. A matrix was used to show the tested knowledge facts in OntoKBCF. According to the knowledge facts, some more hypothetical 'knowledge facts' (bio-health data) would be **assigned** to each test record to make sure most of the types of knowledge facts in OntoKBCF had been used in the test.

I used the following heuristics in constructing the test records: 1) use real parameters as much as possible, 2) use as few test records as possible, 3) represent as many different mutation types as possible and 4) cover the knowledge facts in OntoKBCF as wide as possible.

It should be pointed out that many descriptions were not sufficiently detailed to provide suitable entries for factors in the ontology. For example, "in a 21-year-old black woman with substantial pancreatic disease but only mild pulmonary involvement", "mild pulmonary involvement" is not a detailed clinical description for a record restructuring. For a doctor in practice maybe that the description would be enough to understand the patient's pulmonary status; however, if we want to record the case with clearer and specific detail, that description is far from useful. So the restructured test records are fragmental rather than a set of complete records, and hypothetical data is necessary.

5.6 Mapping between OntoKBCF and the EHR system prototype

OntoKBCF contains basic concepts, combined concepts and final knowledge facts. Basic and combined concepts are used as components to construct and represent the final knowledge fact. All the basic and combined concepts are embodied as hierarchies in OntoKBCF. Final knowledge facts are embodied in their hierarchies (basic and combined concepts) and property descriptions (Asserted Conditions in Protégé-OWL). The basic and combined concepts in OntoKBCF could be used as a data source for the EHR system prototype. The following section describes how mapping is used to make this work.

5.6.1 General introduction

Semantic mapping is carried out manually between OntoKBCF and the CCR labels, based on definitions of the labels in the CCR.

According to how the knowledge facts will be used, the concepts in OntoKBCF can be classified into basic concepts, combined concepts and final knowledge facts; the concepts can also be split into health and biological parts according to the topic. The classification of the knowledge facts in OntoKBCF helps the mapping.

5.6.2 Mapping procedures

These are the mapping procedures used. **First** the major selected labels are listed: Problems, Results, and Procedures. **Secondly** all the defined concepts in OntoKBCF are examined. Then specific labels, which are more specific than major labels, are selected according to the major labels' definitions and examples in the CCR specification (such as 'condition' and 'symptom' which are both specific type of 'Problems') and within the scope of OntoKBCF. **Thirdly** all the defined concepts were reorganised under each specific label through both MS Access tables and structure in Protégé-OWL. The procedures to choose labels from CCR and to map the contents in OntoKBCF to the CCR labels were interactive.

5.6.3 Mapping rules

When all the concepts are reviewed, for the health parts both basic and combined concepts and final knowledge facts were used in mapping; for the biological parts, only

CFTR mutations were used as final knowledge facts. There are two reasons for this choice:

- First, the basic and combined concepts in health part act as a reference for users in EHR-s although they are not a necessary part of the EHR-s.
- Secondly there are many more detailed descriptions about health (clinical) concepts in the CCR specification. These descriptions provide many more mapping points for the health part between OntoKBCF and the EHR system prototype, compared with mapping points for the biological part between OntoKBCF and the EHR system prototype.

There is no specific section or label in the CCR specification to describe biological information. According to the label definition in the CCR specification the CFTR mutations are put under 'Results'. Although there is not any example of mutation (existing examples included haematology, chemistry, virology etc), literally, mutation can be seen as a type of laboratory result. So 'mutation result' was created as a specific label of 'Results' used to collect different CFTR sequence variations. Because there is no corresponding description for biological information in the CCR specification, there is no point in including too much detail biological basic and combined concepts in mapping. All the biological basic concepts, and part of the combined concepts in OntoKBCF are excluded in mapping, such as all the nucleotides, all the amino acids etc.

Some defined temporal, spatial and qualitative concepts and different CF patient groups have not been used in mapping, such as 'Prolonged_disease_course', 'Long_arm_of_chromosomes_human_pair_7', and 'Adolescent_female_CF' etc. Some non self-defined health concepts are used in mapping, for example 'Atresia' is under 'Conditions' which is under 'Problems'; 'Rash' is under 'Symptoms', which is a type of 'Problems'; 'Cystic_fibrosis' is under 'Diagnostic_result' which is a type of 'Results' etc.

5.6.4 Mapping example

There are several mapping examples in the following context. 'Distended_abdomen' is a defined health concept, a description for one of manifestations for infant cystic fibrosis patient and also a combined concept. The concept is used as a property in describing final knowledge facts. In mapping, 'Distended_abdomen' was put under 'Symptoms' which was a type of 'Problems'.

'Delta_F508' is a CFTR mutation and also a combined concept in OntoKBCF. In mapping it was put under 'Mutation result', which was a type of 'Results'. However, I did not map 'F508', which is another combined concept, to describe the amino acid location in the CFTR protein, nor 'Phe', which is a biological concept and a basic concept - Phenylalanine (the abbreviation is F).

'Bronchiectasis' is a basic concept of the health part in OntoKBCF. 'Bronchiectasis' was mapped under 'Diagnosis' which was a type of 'Results'.

5.6.5 Mapping results

The mapping results are a set of tables, about mutational results, diagnostic procedures, conditions, symptoms etc.

The mapping between OntoKBCF and the CCR labels in the EHR system prototype is a test of the content of OntoKBCF. This is an extra use of OntoKBCF as a data source in the EHR system prototype. In an actual EHR system, data sources for EHR such as symptoms or conditions, diagnosis etc, should be provided separately from the EHR system, not necessarily from OntoKBCF.

Although mapping can be achieved in this way, there are some difficulties during this process.

5.6.6 Mapping difficulties

To select proper labels from CCR is not too difficult as there is a clear aim of how to use the EHR system prototype and a clear scope about the content in OntoKBCF. The really difficult part is in mapping, because in the CCR specification the definitions of the labels were not given in detail. There are many design decisions which need to be made, since there is a lack of consensus concerning the definitions of many concepts during mapping. This is easy for some concepts, such as 'coughing', which should be a member of symptoms; but it is more difficult for other concepts, such as: poor weight gain, weight loss. Should they belong to 'conditions', 'symptoms', or both? The decision heavily depends on the definition of the labels. The detailed discussion will be in section 5.9.2.

5.7 Verification of the prototype components

The research seeks to show the feasibility of bringing biological information into a simulated clinical system by integrating a knowledge base prototype with an EHR system prototype. The research requires the development of two components, i.e. the substantive knowledge base- OntoKBCF and the EHR system prototype. In addition the means of connecting the two were investigated, i.e., how the connection shows the structure and relationships specified by the ontology and enable the knowledge base to be tested. Furthermore, the connection demonstrates the feasibility of bringing the two different types of information together within a prototype of an EHR system.

In this section we evaluate the research. The background introduction about evaluation has been presented in section 3.5. This evaluation does not include the evaluation of the EHR system prototype per se, which is simply a means of showing the knowledge base's structure and content. There are functional criteria for full EHR systems and some of those criteria have been used in building the prototype. However, the main purpose of the prototype is to embody the CCR specification to which the knowledge base attempts to map. In this section only connection-related parts of the EHR will be evaluated; for example, the biological information is made accessible through the EHR prototype, and filters are there to help to make the information relevant etc.

There are different types and levels of testing during the software development process[176]. Two terms are often used in software testing: verification and validation[176, 177]; verification concerns the correctness of a specification whereas validation is checking a piece of software against the requirements, which is related to the process of executing test data in the software[177]. Evaluation in this thesis is concerned with verification. This is a functional verification rather than a usability evaluation. Screen shots are used to show the functional verification results and test records with scenarios are used to make the examples explicit. The method of defining the test records has already been described in section 5.5.2.

There are two major parts of verification in this section, i.e., OntoKBCF and the connection between OntoKBCF and the EHR prototype.

- Evaluation of OntoKBCF concerns structure and content of the knowledge base prototype. The objectives of OntoKBCF's evaluation are 1) to prove that the logical structure of OntoKBCF is consistent and correct; and 2) the content of OntoKBCF is useful and usable, by showing the mapping with CCR headings. This is dealt with in section 5.7.1.

- The functional verification of the connection part comprises 1) the means of access via the EHR prototype; 2) the use of filters to make the EHR system more active; 3) the presentation of the knowledge. This is dealt with in sections 5.7.2-5.7.4.

5.7.1 Evaluation of OntoKBCF

Evaluating ontologies is a complex task. However, because ontologies are logical specifications, much work has been done to provide automated tools to test the consistency and correctness of an ontology. These tools are called 'reasoners' A reasoner is "a piece of software able to infer logical consequences from a set of asserted facts", which can help to keep consistency and explicit hierarchy of the ontology[8, 178]. RacerPro[179] is a description logic reasoner and is the appropriate tool for use here as this project uses the OWL-DL in OntoKBCF construction.

Two methods were then used to evaluate the ontology:

1) A formal reasoner, RacerPro 1.9.0, was used to:

   - check the consistency of OntoKBCF; OntoKBCF passed the reasoning process without fault in 5.789 seconds (there were 405 named classes and 208 anonymous classes, and 112 named classes had been defined;);
   - check the classification of OntoKBCF took 57.993 seconds to complete.

2) Another method used relates to the successful transfer of OntoKBCF across different tools, i.e., Protégé-OWL (version 3.3.1) and TopBraid (test version 2.3.1). This interchange also succeeded and provided further evidence for the consistency and validity of OntoKBCF's structure.

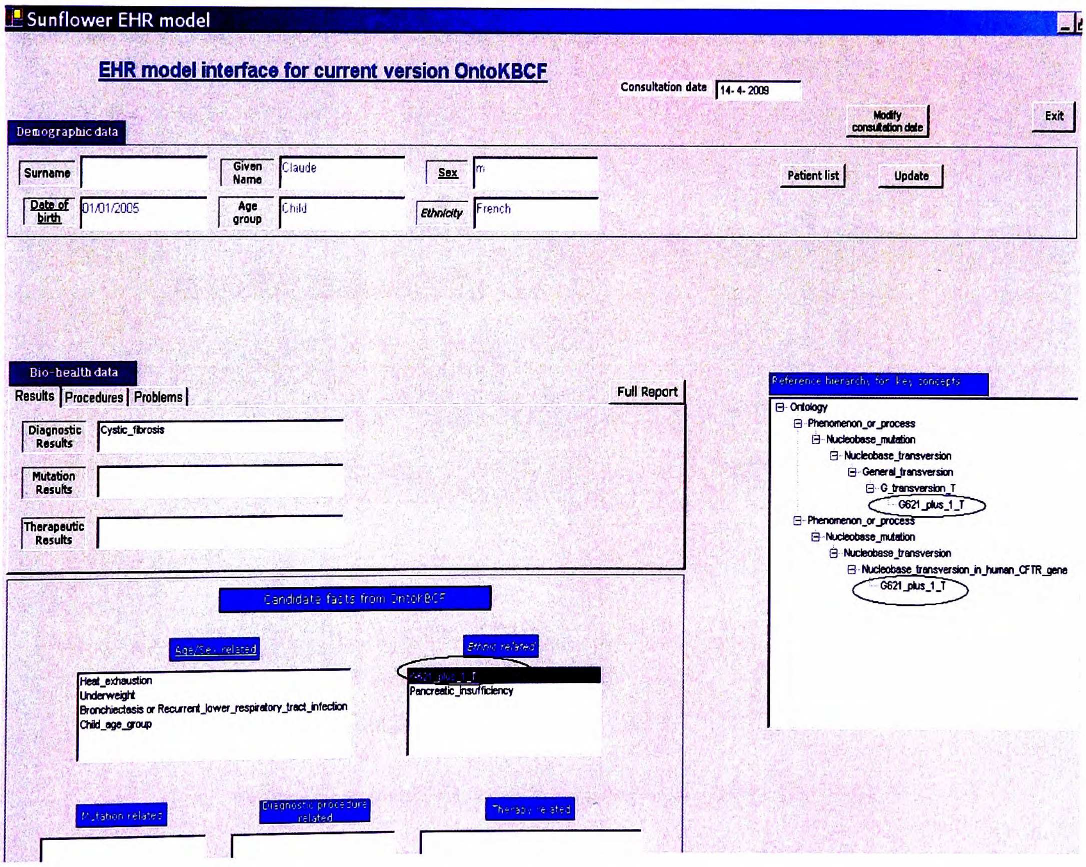
Mapping between the CCR labels and OntoKBCF was an evaluation related to the appropriateness of the contents in OntoKBCF. All the mapping between the CCR labels and OntoKBCF contributes to the bio-health candidate tables. Furthermore, all the bio-health data used in the testing of the connection was from these tables. This means that the content in OntoKBCF was suitable for our purpose. Other positives for the structure of OntoKBCF, such as reduced redundancy, and a robust and explicit structure were simply gained from direct experience in ontology restructuring, mapping, and modification during the construction process.

5.7.2 Means of access via the EHR Prototype

The bio-health information in OntoKBCF is made accessible through the EHR prototype. The following scenario is used to demonstrate this function.

> Claude, male, 01/01/2005, French; he has cystic fibrosis.
>
> In this scenario the ***candidate*** facts presented via the EHR's interface from OntoKBCF include: 'patient could have heat exhaustion', 'be underweight' and 'have bronchiectasis or lower respiratory tract infection and pancreatic insufficiency'; 'possible CFTR mutation type is G621+1T'. The tree view consists of hierarchies of G621+1T (Figure 5-7).

The means of access via the EHR prototype is embodied in five OntoKBCF candidate facts boxes and a tree view structure; both types of structure are from OntoKBCF (see hierarchy and candidate facts, Figure 5-7). G621+1T (candidate mutation) is a representative of biological information which is introduced into the EHR record.



Consultation date: 14- 4- 2009

Surname:
Given Name: Claude
Sex: m
Date of birth: 01/01/2005
Age group: Child
Ethnicity: French

Diagnostic Results: Cystic_fibrosis
Mutation Results:
Therapeutic Results:

Age/Sex related:
Heat_exhaustion
Underweight
Bronchiectasis or Recurrent_lower_respiratory_tract_infection
Child_age_group

Ethnic related:
G621_plus_1_T
Pancreatic_insufficiency

Reference hierarchy for key concepts:

```
⊟ Ontology
   ⊟ Phenomenon_or_process
      ⊟ Nucleobase_mutation
         ⊟ Nucleobase_transversion
            ⊟ General_transversion
               ⊟ G_transversion_T
                     G621_plus_1_T
   ⊟ Phenomenon_or_process
      ⊟ Nucleobase_mutation
         ⊟ Nucleobase_transversion
            ⊟ Nucleobase_transversion_in_human_CFTR_gene
                  G621_plus_1_T
```

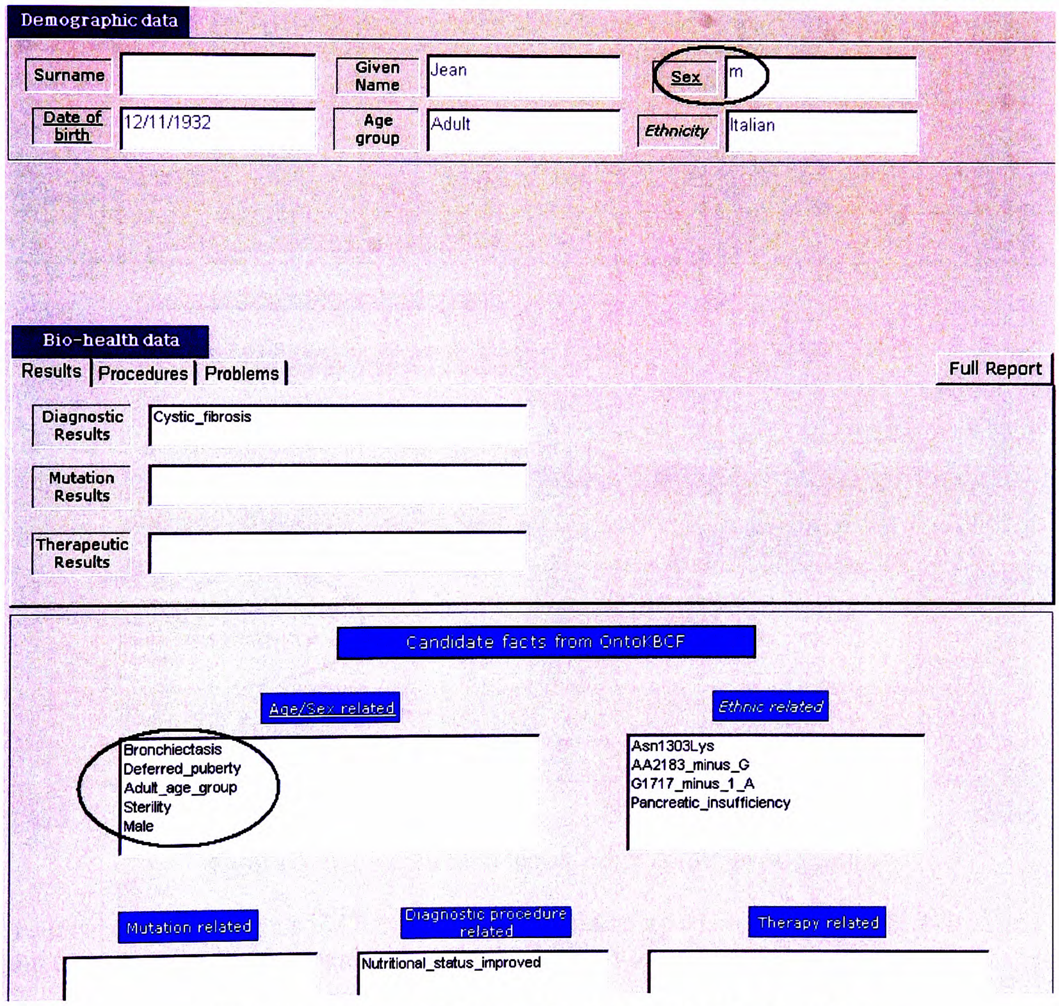
Figure 5-7 Screenshot for accessing bio-health information in OntoKBCF through the EHR prototype

(Note: general EHR interface including demographic and bio-health data, candidate knowledge facts from OntoKBCF and tree structure of G621+1T)

5.7.3 Use of filters to make the EHR-s more active

In the connection section the bio-health information in OntoKBCF can be customised according to individual patients' parameters, i.e. the EHR prototype could be set to display personalised bio-health information according to the filters, which include demographic data and bio-health data. The following scenario is used to show this function:

> Jean, male, 12/11/1932, Italian; he has cystic fibrosis.
>
> In this scenario the ***candidate*** facts on the EHR interface from OntoKBCF include: 'that patient might have bronchiectasis', 'deferred puberty', 'sterility'; and 'pancreatic insufficiency'; related mutations include Asn1303Lys, AA2183-G, G1717-1A (Figure 5-8).



Demographic data

Surname		Given Name	Jean	Sex	m
Date of birth	12/11/1932	Age group	Adult	Ethnicity	Italian

Bio-health data

Diagnostic Results	Cystic_fibrosis
Mutation Results	
Therapeutic Results	

Candidate facts from OntoKBCF

Age/Sex related:
Bronchiectasis
Deferred_puberty
Adult_age_group
Sterility
Male

Ethnic related:
Asn1303Lys
AA2183_minus_G
G1717_minus_1_A
Pancreatic_insufficiency

Mutation related:

Diagnostic procedure related:
Nutritional_status_improved

Therapy related:

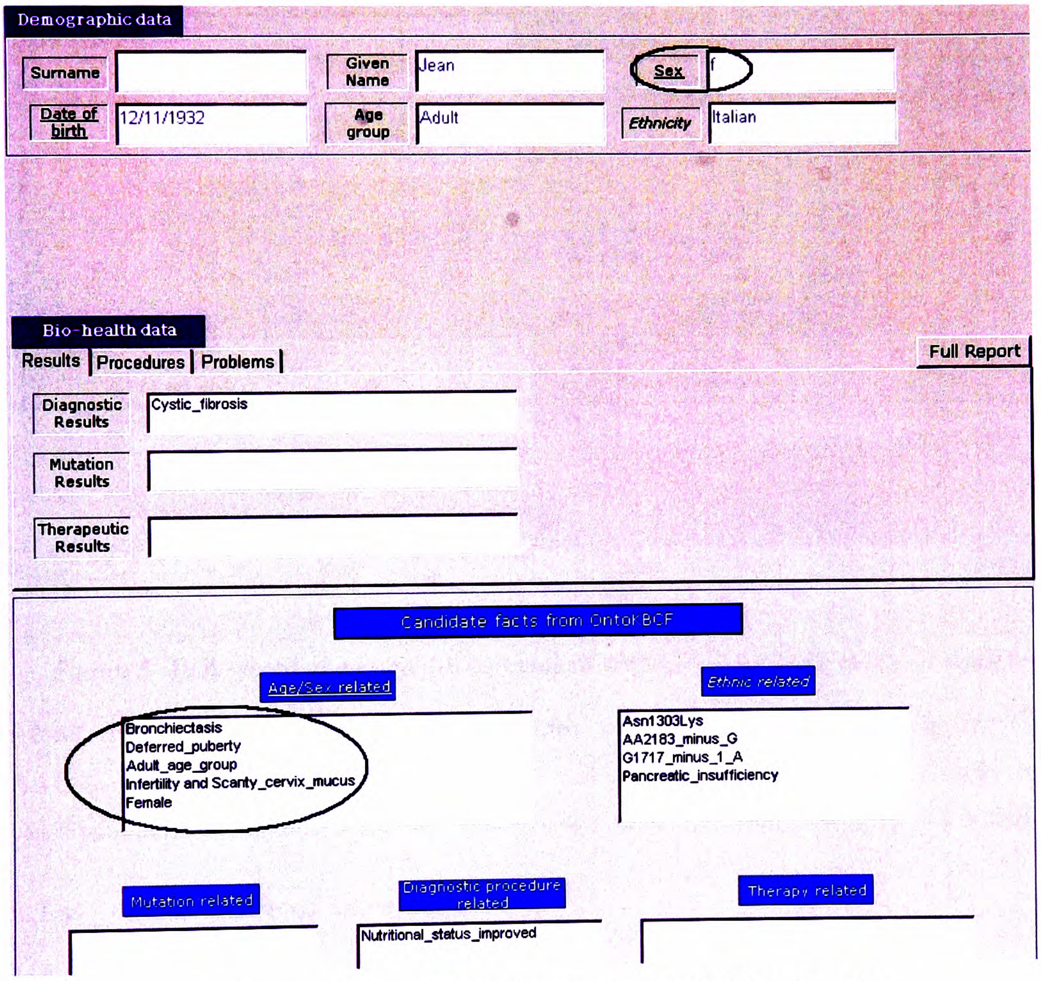
Figure 5- 8 A record for a male adult cystic fibrosis patient

(Note: the sex is related with the candidate facts displayed in the Age/Sex box -lower circle- and it can be compared with Figure 5-9, which is the female version of the same patient)

Figure 5-8 is the record for male Jean. Figure 5-9 is a record for female Jean (i.e. as a female version but all other parameters are the same with male version except for sex).

The Age/Sex related box shows different candidate facts in Figure 5-8 from in Figure 5-9 because of the different sex. This is to show 'sex' can be set as a filter. Sex is a representative of demographic data.

Figure 5-10 shows that if a patient's record has an AA2183-G then AA2183-G will not appear in the candidate fact box. This shows bio-health data as a filter. The mutation is a representative filter for bio-health data.



Surname	
Given Name	Jean
Sex	f
Date of birth	12/11/1932
Age group	Adult
Ethnicity	Italian
Diagnostic Results	Cystic_fibrosis
Mutation Results	
Therapeutic Results	

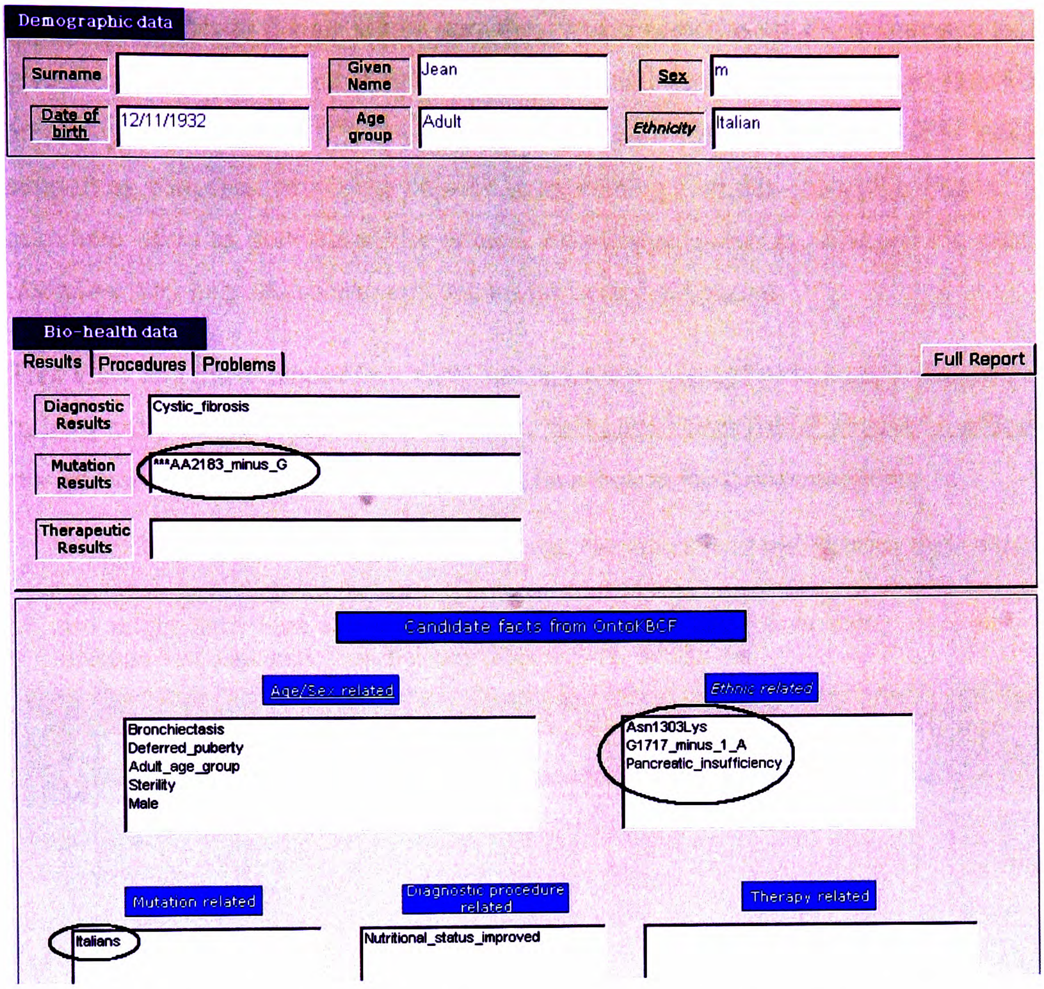
Bronchiectasis
Deferred_puberty
Adult_age_group
Infertility and Scanty_cervix_mucus
Female

Asn1303Lys
AA2183_minus_G
G1717_minus_1_A
Pancreatic_insufficiency

Nutritional_status_improved

Figure 5- 9 A record for a female adult cystic fibrosis patient

(Note: except for sex- this is a female, other parameters are the same as the patient in Figure 5-8 and correspondingly the candidate facts in the Age/Sex box are different - lower circle)



Surname		Given Name	Jean	Sex	m
Date of birth	12/11/1932	Age group	Adult	Ethnicity	Italian

Diagnostic Results	Cystic_fibrosis
Mutation Results	***AA2183_minus_G
Therapeutic Results	

Age/Sex related	Bronchiectasis Deferred_puberty Adult_age_group Sterility Male
Ethnic related	Asn1303Lys G1717_minus_1_A Pancreatic_insufficiency
Mutation related	Italians
Diagnostic procedure related	Nutritional_status_improved
Therapy related	

Figure 5- 10 A record of a cystic fibrosis patient with CFTR AA2183_minus_G mutation

(Note: this is the same patient as in Figure 5-8 except that this patient has CFTR AA2183_minus_G already- upper circle-, so AA2183_minus_G would not display in the candidate facts- middle circle- and this mutation may relate to an Italian patient- lower circle)

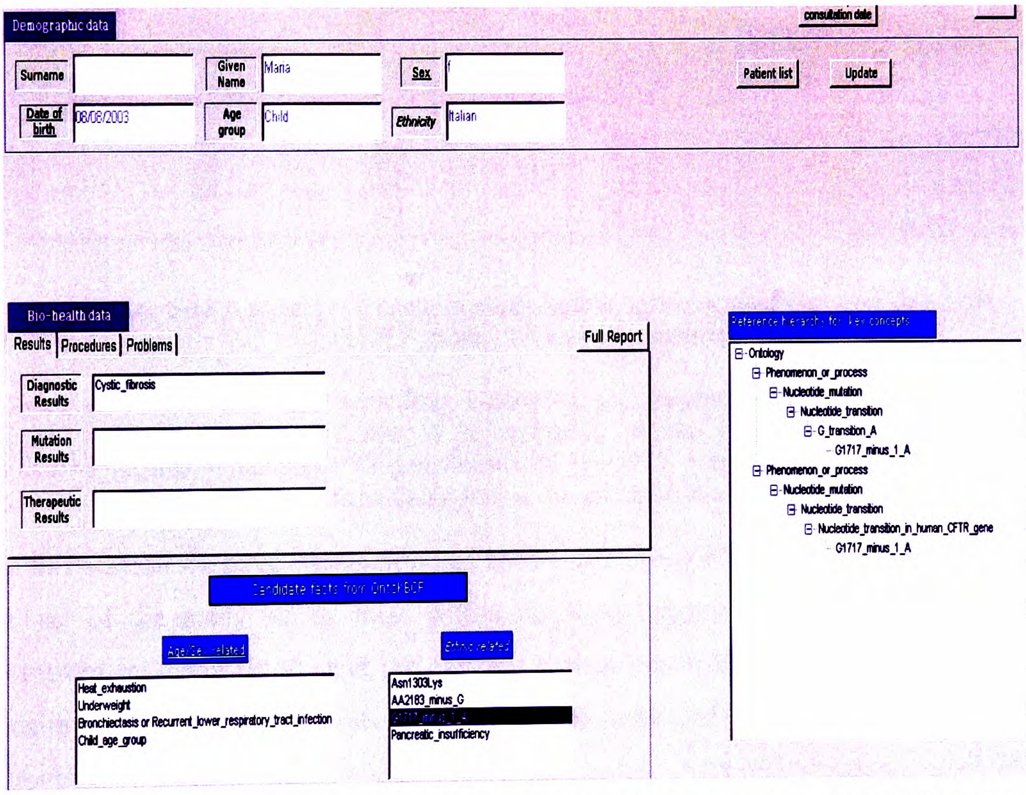
5.7.4 Presentation of knowledge

This section shows how the connection part can present bio-health knowledge through the EHR prototype. In this section we include the test patient record interface with 1) tree view (Figure 5-11, for 'G1717_minus_1_A' mutation); 2) extended tree view and 3) natural language interpretation with a confirmation dialogue before the candidate fact is added to the existing bio-health data (Figure 5-12, for 'G1717_minus_1_A' mutation to be added to 'mutation results').

Two different ways of presenting the knowledge are shown as examples. Eventually users will be able to decide which way they would prefer to see the knowledge so as to give them the most support. Knowledge is presented through 1) a tree view (knowledge hierarchy from OntoKBCF); 2) an extended tree view (knowledge hierarchies including supporting concepts, which can be used in explaining complex concepts). The tree view may help users to understand the domain knowledge hierarchy, whereas the extended tree view may help the user to understand the complex concept.

Tree view and the extended tree view can be used as general or more specific references to help users decide whether the candidate facts from OntoKBCF should be added into the patient's record (the bio-health data). The scenario for this example is:

> Maria, female, 08/08/2003, has cystic fibrosis, meconium ileus and receives physiotherapy.
>
> Candidate facts include: patient might have Asn1303Lys, AA2183-G or G1717-1A; patient also might have heat exhaustion, underweight, bronchiectasis or lower respiratory tract infection and pancreatic insufficiency (Figure 5-11, 5-12).


Demographic data

Surname:
Given Name: Maria
Sex: f
Date of birth: 08/08/2003
Age group: Child
Ethnicity: Italian

Bio-health data

Diagnostic Results: Cystic_fibrosis
Mutation Results:
Therapeutic Results:

Candidate facts from OntoKBCF

Age/Sex related:
Heat_exhaustion
Underweight
Bronchiectasis or Recurrent_lower_respiratory_tract_infection
Child_age_group

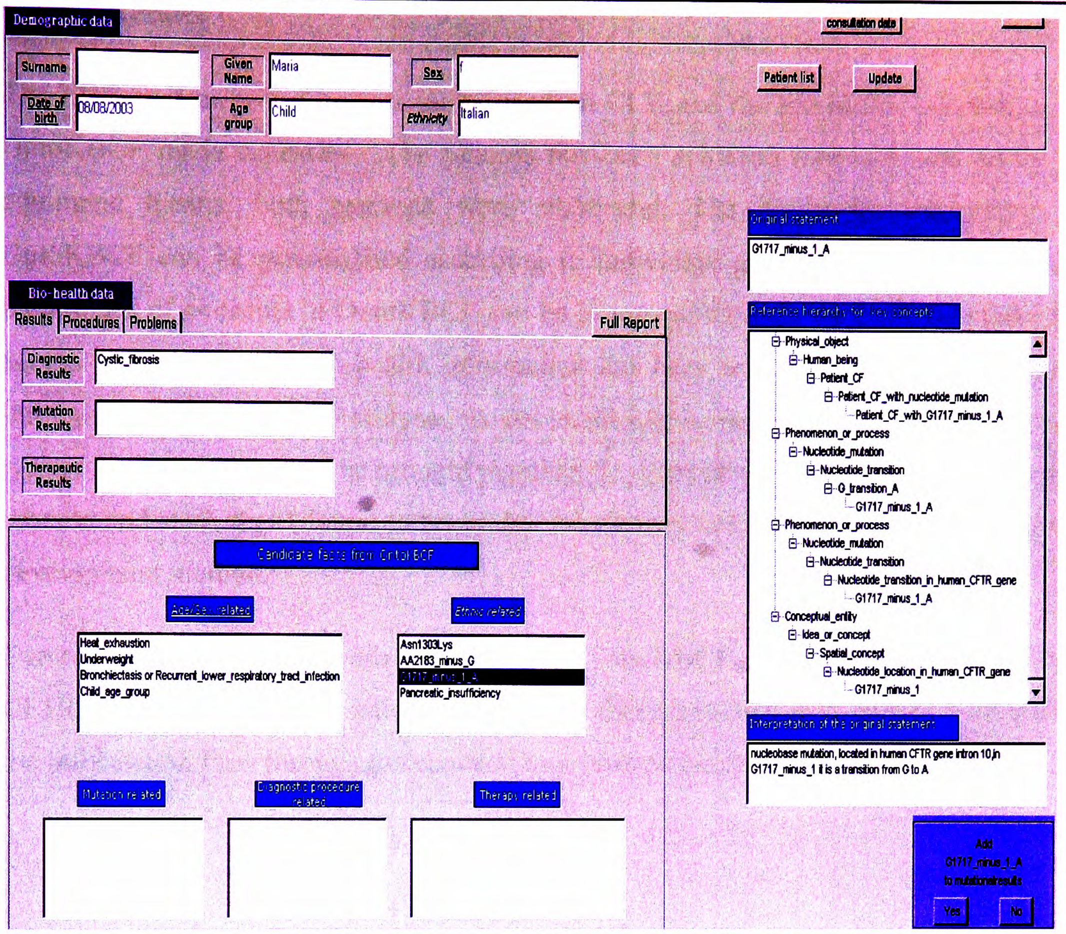
Ethnic related:
Asn1303Lys
AA2183_minus_G
G1717_minus_1_A
Pancreatic_insufficiency

Reference hierarchy for key concepts

```
⊟ Ontology
  ⊟ Phenomenon_or_process
    ⊟ Nucleotide_mutation
      ⊟ Nucleotide_transition
        ⊟ G_transition_A
          - G1717_minus_1_A
  ⊟ Phenomenon_or_process
    ⊟ Nucleotide_mutation
      ⊟ Nucleotide_transition
        ⊟ Nucleotide_transition_in_human_CFTR_gene
          - G1717_minus_1_A
```

Figure 5- 11 A record of a cystic fibrosis patient with basic tree view of CFTR G1717-1A

(Note: basic tree view can provide knowledge hierarchies of CFTR G1717-1A)



Surname:
Given Name: Maria
Sex: f
Date of birth: 08/08/2003
Age group: Child
Ethnicity: Italian

Diagnostic Results: Cystic_fibrosis
Mutation Results:
Therapeutic Results:

Age/Sex related:
Heat_exhaustion
Underweight
Bronchiectasis or Recurrent_lower_respiratory_tract_infection
Child_age_group

Ethnic related:
Asn1303Lys
AA2183_minus_G
G1717_minus_1_A
Pancreatic_insufficiency

Original statement: G1717_minus_1_A

Reference hierarchy for key concepts:

```
⊟ Physical_object
    ⊟ Human_being
        ⊟ Patient_CF
            ⊟ Patient_CF_with_nucleotide_mutation
                Patient_CF_with_G1717_minus_1_A
⊟ Phenomenon_or_process
    ⊟ Nucleotide_mutation
        ⊟ Nucleotide_transition
            ⊟ G_transition_A
                G1717_minus_1_A
⊟ Phenomenon_or_process
    ⊟ Nucleotide_mutation
        ⊟ Nucleotide_transition
            ⊟ Nucleotide_transition_in_human_CFTR_gene
                G1717_minus_1_A
⊟ Conceptual_entity
    ⊟ Idea_or_concept
        ⊟ Spatial_concept
            ⊟ Nucleotide_location_in_human_CFTR_gene
                G1717_minus_1
```

Interpretation of the original statement: nucleobase mutation, located in human CFTR gene intron 10,in G1717_minus_1 it is a transition from G to A

Add G1717_minus_1_A to mutationalresults

Figure 5-12 A record of a cystic fibrosis patient with extended tree view of CFTR G1717_minus_1A and interpretation

(Note: extended tree view including hierarchies of Patient_CF_with_G1717_minus_1_A, G1717_minus_1_A, G1717_minus_1; Interpretation of the original statement "nucleotide mutation, located in human CFTR gene intron 10, in G1717_minus_1 it is a transition from G to A". The user may use the confirm dialog before the mutation is added to mutation results)

In this section we have shown how the knowledge base delivers candidate facts; a small subset of the many stored facts within the knowledgebase. The candidate facts are retrieved by using filters, and the existing patient health information stored within the health record. Retrieval and presentation are triggered automatically according to submitted patient details, making the record system more of an active rather than a passive system.

Although the research aims have been achieved, there are limitations and there are also future directions that need to be explored; these will be summarised in section 5.9.

5.8 Results

The EHR system prototype was built based on CCR and the bio-health information in OntoKBCF made accessible. The process has been achieved manually and partly by automatic means; both methods were successful. The bio-health information in OntoKBCF can be personalised according to individual patients' parameters and the hierarchies of concepts in OntoKBCF can be presented through the EHR interface; the accessible personalised bio-health information has been tested by several test records through the EHR system prototype. At the beginning it was unclear as to how the two types of information could be brought together. To close this results section I include the process and steps that I found to be useful, which may help others to produce a formal development method.

Currently filters for bio-health information in OntoKBCF include: age, sex, ethnicity, CFTR mutation type, diagnostic procedure and therapy related data. Figure 5-13 shows the information flow through the connection prototype interface.

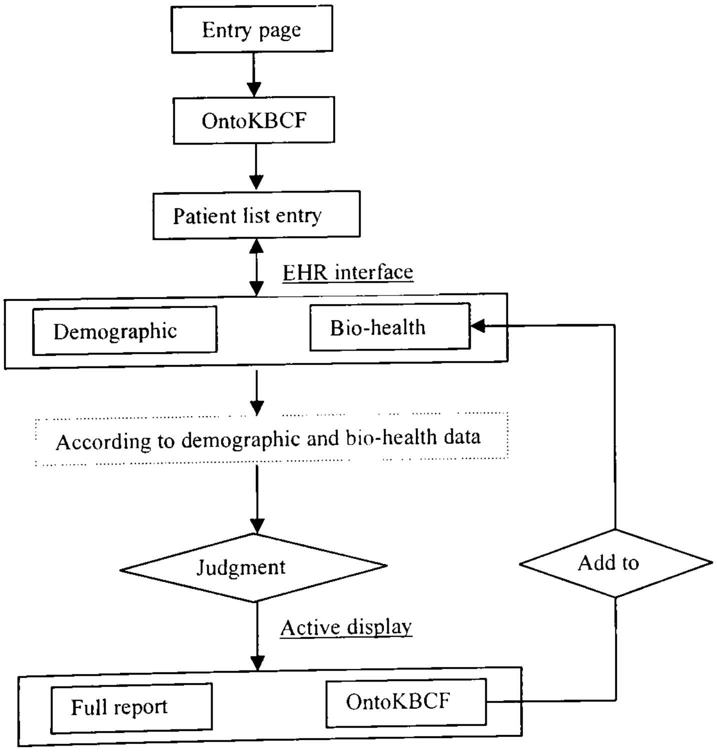


Figure 5-13 Information flow diagram for the EHR/OntoKBCF connection prototype

(Note: double direction arrow means read and write, which allows edit function)

The results of the automatically derived part are a set of MS Access tables, which correspond to the Access tables produced manually; however, the two sets are not identical. The tables or part of the tables involved with hierarchies of OntoKBCF could be created totally automatically; the tables or part of the tables involved with property description or logic representation could not be created automatically, because these parts of content in OntoKBCF could not be displayed correctly in the original converted table from Protégé-OWL. More details are given in section 5.9.3.

This research has been a "learning through making" process. On reflection, the process and steps that I used to connect a knowledge base prototype with an EHR system prototype were:

1) To decide what to offer and how to offer it from the knowledge base prototype through the EHR prototype interface.

2) To construct a compatible EHR structure in the knowledge base prototype according to the requirements set in the first step. There were several specific points to be considered:

    a) What structure in the knowledge base prototype need to be displayed or accessed through the EHR prototype: hierarchy only, or both hierarchy and property?

    b) If hierarchy only, then should the original hierarchy only be used or the classes organised in a different way?

    c) The EHR structure in knowledge base prototype need to be compatible with the EHR prototype architecture.

3) To make a technical choice, which type of knowledge base prototype file should be used: converted table, XML, or RDF file?

4) To organise the interpreted content from the knowledge base prototype manually or automatically according to what would be offered through the EHR prototype.

5) To program for both EHR prototype and the knowledge base prototype file to make the connection work.

5.9 Discussion

Although as a proof of concept research, the research objectives were achieved successfully: i.e., to make the formal and structured bio-health information accessible and personalised through a CCR-based EHR system prototype. There were, however, many limitations and problems with the connection process pointing to future directions that remain to be explored.

5.9.1 CCR limitations

CCR was used in building the EHR system prototype; however there is no biological specification in CCR. The clinical specifications are not detailed in CCR, so many design decisions had to be made in choosing labels and in the mapping process. In the following context several examples will be used to describe the problems faced in using CCR.

'Laboratory results' and 'therapeutic results' are two examples in the definition of 'Results' (i.e. one of the CCR attributes). However, there is no further explanation about how to distinguish the two labels. Semantically, therapeutic results and laboratory results can overlap with each other since therapeutic results can be indicated by laboratory results. It would be very difficult to decide which label should be chosen without detailed differentiation between the two labels. For example 'Increased_value_of_mucus_clearance', 'Bone_density_increased', 'Nutritional_status_improved': should these belong to 'laboratory results', 'therapeutic results', or both? If there is an exact value for bone density, then the specific value of the bone density should be collected under the laboratory result. However, in most reviews of patient's records, or extracted patient's records, there is a lack of detail about the examination that the patient has taken. Currently all the three examples above are collected under "therapeutic results", however, this has been a design decision only.

'Conditions' and 'diagnoses' are two examples for 'Problems' (a CCR attribute) while 'diagnostic results' is an example for 'Results' (another CCR attribute). How should the three labels be distinguished? What types of relationships are there among the three labels? This is unclear. Under 'Problems' I did not include 'diagnoses', but I did use 'diagnostic results' under 'Results'.

In the current EHR system prototype we have not included different mutation test methods. According to the CCR specification, different mutation test methods can be included in 'diagnostic procedure'. However, the mutation results will be in 'laboratory results'. Considering there is a heading- 'diagnostic results' in CCR, this would be inconsistent if details of different mutation test methods were included. Which should be the heading for mutation results: 'laboratory results' or 'diagnostic results'? It is not the intention of this research to provide an exclusive CFTR mutation, but rather a representative of CFTR mutation.

Generally, there is a lack of biological labels and a lack of sufficiently detailed difference/explanation/differentiation among the various labels in CCR.

5.9.2 Mapping limitations and merits

Mapping between OntoKBCF and the CCR labels manually can work well. However, in the automation part, there are some limitations.

How to keep the mapping reasonable without sacrificing simplicity in the coding was a major challenge. Logically 'reasonable' means placing both parents and children under the same CCR label as parents and children. For example: 'Nasal_polyps', and 'Recurrent_nasal_polyps': semantically both of the concepts should be collected under 'Symptoms' The actual hierarchy in the domain (KBCF) part of OntoKBCF is that 'Recurrent_nasal_polyps' is a child (subclass) of 'Nasal_polyps. However, the class and subclass relationship will be too complex for coding should we want to make the procedure automatic. To define the two concepts as siblings under the concept 'Symptoms' would be much easier for coding, but it would be a logical conflict. The hierarchical relationship in the domain part is parent and child and they cannot be siblings in the EHR part within the same knowledge base. Only one of the concepts was kept in the EHR part in the real practice. This was a pragmatic decision.

In the current version mapping focuses mainly on hierarchies of concepts, without considering the 'property' descriptor. Mapping simply based on the hierarchy only is not sufficiently detailed to represent the exact meaning, and hence there will be some information loss. 'Recombinant_human_deoxyribonuclease' is a subclass of "substance" in the domain hierarchy; 'treated_by' (property) is used to represent the therapy using this substance in OntoKBCF. There is not a method listed as a separate subclass of therapy in OntoKBCF. The hierarchy-only method cannot express the original meaning completely and precisely.

Although there are some limitations in mapping, the mapping structure in OntoKBCF brought some advantages for knowledge representation in the domain part as well. The EHR mapping structure provided help to OntoKBCF with more constraints in knowledge representation. For example, in the property 'has_manifestation_of', with EHR structure in OntoKBCF, we can restrict the property in 'symptoms' and 'conditions' classes. Before the EHR structure was added into OntoKBCF, the subclasses of the two classes were distributed in every corner of OntoKBCF; it was not realistic to restrict the property in every single class.

5.9.3 Automation limitations

It would be ideal to make the connection part completely automatic, which is critical for future updating; however there are still limitations to this process.

The tables or the part of the tables in the Access database could not be extracted automatically if the required fields were related to the property description, for example "has_manifestation some Pancreatic_insufficiency", or "occur_in_age_group some Adolescent_age_group", because the representation could not be displayed correctly in the converted table. The content would be displayed as an internal string, such as "@_:A98". So the tables used in the EHR prototype could not be created completely automatically from the converted table.

5.9.4 Nomenclature of the titles in the EHR interface

The titles of the five OntoKBCF-related boxes in the EHR interface are related to their origin, not necessarily related to the contents/items themselves listed in the boxes. For example, in the ethnic related box, any knowledge fact related to particular ethnicity (such as the most possible CFTR mutation related to Swiss or Italian patients) will be listed, however, the knowledge facts themselves are not about ethnicity (in this case they were caused by CFTR mutations).

In one of OntoKBCF-related boxes' titles in the EHR system prototype interface, 'Therapy related' is used instead of 'Therapeutic procedures related' or 'Therapeutic results related' This is because the knowledge facts themselves, which are listed in the box, are therapeutic results, which result from the corresponding therapeutic procedures. This means the facts in this box are related to both 'Therapeutic procedures' and

'Therapeutic results'. So 'Therapy related' was therefore chosen as the box title.

5.10 Summary

In this chapter I have presented the EHR system prototype interface, its construction and its connection with OntoKBCF; what function the EHR system prototype and the connection part have; how to automate the connection; how to map OntoKBCF and the CCR labels; what are the current results; and what are the known limitations and difficulties. Generally I have answered questions about "what", "how", and "why" for the connection between OntoKBCF and the EHR system prototype in this chapter. This is an extension for OntoKBCF in both content structure and application. This is also an exploration towards a more active future EHR system prototype.

Chapter 6 Discussion

The ability to deliver formal and structured bio-health information in a clinical setting has huge potential. It is key to personalised health care services, since biological information is fundamental to understanding more of the unique traits of an individual and also plays a critical role in the determination and explanation of disease. In this research such personalised information has been presented via a CCR-based EHR interface and tested by a number of test records. This chapter will restate the research results and interpret them, then compare this research with other research, to present the significance, limitations, experiences and future directions of this research.

6.1 Restatement of results

The major results of this research have been to show:

1) How biological information associated with health information can be organised in a formal, structured way in an ontology-based knowledge base.
2) How the bio-health information can be made accessible through a CCR-based EHR prototype, which not only reveals the content and structure of the underlying knowledge base but presents the personalised bio-health information via the interface.

It is a step towards seamlessly linking the two types of information, which presently remain largely unconnected, but which will inevitably come closer together.

6.2 Interpretation of results

The primary results of this research can be divided into three parts, results of the ontology work - OntoKBCF, the EHR prototype, and the connection between the two. The three parts will be interpreted in turn.

6.2.1 Findings related to OntoKBCF

OntoKBCF has three characteristics:

1) The bio-health information is organised in a formal and structured way.
2) It has both basic concepts and final knowledge facts.
3) It can be used not only as a knowledge management tool but also an

application of semantic web technology.

The bio-health information is organised according to accepted ontological principles found in the literature. The proposed organisation should ideally be capable of improving how such information can be used, reused, shared, and communicated. At the same time the organisation should attempt to decrease information overload to the body of users who already complain of being overwhelmed. In such a short study it is not possible to directly prove all these sub-goals. Therefore part of the strategy was to deploy technologies and develop tools that are already recognised as having some of these benefits. For example, 'reuse', 'sharing' and 'communication' are promises or realisations of existing semantic web technology, and consequently these are also found in OntoKBCF.

Decreasing information overload is one sub-goal that is more difficult to show in a study like this, because 'overload' might be considered to be a subjective concept, depending on the user and their context at any one point in time, and user-testing is beyond the scope of this study. However, it is possible to organise the knowledge base model in a principled way that should make the sub-goals achievable:

1) The fact that the bio-health information in OntoKBCF is based on an ontology permits a more systematic organisation, and better structure with less redundancy.
2) The explicit relationships within the knowledge base permit information to be tailored to a defined context rather than delivering the whole set of information in one go. The association of biological and health information in OntoKBCF makes it feasible for further mapping and information customization via the EHR prototype interface.
3) It provides basic concepts and final knowledge facts from the same source.

There is an argument that by bringing bio and health information together automatically an increased volume of information will be created and therefore 'overload' is inevitable. However, the inescapable fact that these types of information are required and will be brought together means that tools such as OntoKBCF and the approach used here may be used in the future to either minimise the excess potential or perhaps through sophisticated filters and tailoring to avoid deluge. This research has started this work.

An original aspect of this research is to attempt to avoid knowledge overload by precise knowledge extraction because of clinically-driven queries.

From the structural point of view of OntoKBCF, there are basic concepts (including both biological and health concepts, for example '**Gly**', which is Glycine), combined concepts (i.e., more specific classes based on existing classes from UMLS or GO, for example '***Gly542X***', which is an amino acid substitution), and final knowledge facts (i.e., concepts with associated OWL conditions, for example the properties related to '***Patient_CF_with_Gly542X***'). This structure has provided a stable foundation throughout the project. It permits restructuring and revision without compromising the overall integrity of the ontology. It also proves useful in establishing connections between the knowledge base prototype and the EHR prototype.

OntoKBCF uses different cystic fibrosis patient groups (such as patients in different age and patients with different CFTR mutations etc) as subjects to organise the related properties together. The knowledge facts organised by this method can be properly accessed through the EHR prototype (although only by manual processes at this time).

The structure provides a useful way of navigating content. For example, a combined concept can be traced down to the appropriate more basic concepts. Although this may not be revealed to doctors completely via the EHR prototype, it is an inexpensive way of showing the relationships between concepts and this might help a user in understanding or interpreting the complex combined concepts. For biological information, the hierarchies of supporting concepts (both basic concepts and combined concepts) for cystic fibrosis patients with different mutations can be made accessible through the EHR prototype. At present supporting concepts for mutations within the knowledge base include the position and mutation type of amino acids or nucleotides; these concepts can be traced by hierarchies.

OntoKBCF is also an application of semantic web technology. It is quite common to use ontological principle in knowledge management, for example in biomedical vocabularies e.g. UMLS and GO, and in knowledge resources e.g. the Foundational Model of Anatomy (FMA). From a functional point of view[180] OntoKBCF can be used as an encyclopaedic knowledge resource for bio-health aspects of cystic fibrosis. The structure and content are accessible through the EHR interface. To use OntoKBCF through such an EHR interface is a step forward in exploring ontology application in a clinical setting.

There are many aspects that could be used to describe a mutation, for example mutation locations (amino acid location, nucleotide location, chromosome location, affected codons etc), mutation types (deletion, insertion, substitution, transition, transversion, complex mutation etc), exact nucleotide or amino acid change, related phenotype (molecular, tissue or organism level). Considering this research targets the non-specialist doctor, such as a GP, who is not a molecular geneticist, only the most basic characteristics of mutation have been included rather than the comprehensive set of characteristics of mutation and interpretation recommendations[71].

6.2.2 Comparison of OntoKBCF with similar research

Existing research is reviewed in Chapter 2 but here I will focus more on the differences between the published research and my solution in OntoKBCF.

PharmGKB[181] and PheGe[53] involved both genotype and phenotype data. In both cases the development work provided a more comprehensive representation at a specific level of hierarchy, i.e. horizontally (Figure 6-1). Compared with those two studies, OntoKBCF provides a representation of bio-health information, starting from basic concepts to final knowledge facts, including both sequence variations information and clinical symptoms that move between levels of the hierarchy, i.e. vertically (Figure 6-1). Full information at any level has not been represented completely in OntoKBCF (Figure 6-1), for example sequence variations information has included some CFTR mutation types without exhaustive CFTR gene mutations or mutation types. Whereas in OntoKBCF both biological and health information is defined in a formal and structured way using ontological principles. By contrast, PharmGKB developed only an XML format for genotype data. OntoKBCF also extends the application to a CCR-based EHR prototype whereas PharmGKB is a stand-alone knowledge base and PheGe is a platform rather than an end-user application for specific content.

CBO[51] was an ontology for describing clinically significant genomics concepts, which involved more detailed vocabulary related to molecular diagnosis and cytogenetics. However, this was confined to the horizontal level and it was very comprehensive. If the content represented in CBO can be described as a surface, then the content represented in OntoKBCF would represent a line which goes through the CBO surface. OntoKBCF provides vocabulary, semantic hierarchies and final knowledge facts. Most of the content in OntoKBCF can be presented in response to patient's parameters being entered through the EHR interface.

Unlike Kumar's research[52], OntoKBCF includes the concepts from nucleotides to disease phenotypes. The content in OntoKBCF is organised on two axes: time-oriented and problem-oriented, and the knowledge scope is established at the outset. However, that constructing a knowledge base is only the first part of this research.

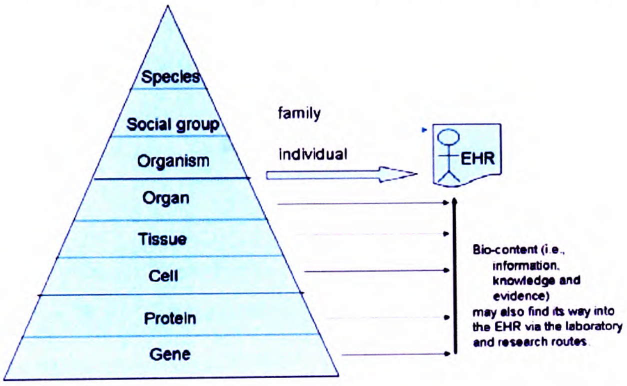

Species
Social group
Organism
Organ
Tissue
Cell
Protein
Gene
family
Individual
EHR
Bio-content (i.e., information, knowledge and evidence) may also find its way into the EHR via the laboratory and research routes.

Figure 6- 1 The EHR as container of diverse and complex data

Compared with GO and UMLS[56, 182], OntoKBCF includes both basic concepts and representation of final knowledge facts. Both of these tools have been used in OntoKBCF's construction.

Unlike LSDB[73, 183], OntoKBCF is based on semantic web technology whereas most LSDB are based on database technology and are stand-alone, for example CFMDB. The formal and structured bio-health information in OntoKBCF can be made accessible through the EHR prototype, which provides a simulated clinical system for OntoKBCF's clinical application.

6.2.3 Findings related to the EHR prototype

The EHR system prototype has several characteristics:

1) Formal and structured biological information can be made accessible.

2) The EHR system prototype is based on an international standard -CCR.

3) Accessible information from OntoKBCF can be filtered according to a patient's

parameters.

4) The hierarchies within the knowledgebase provide a descriptive model of the domain, making explicit and clear the meaning and interrelationships between concepts at the interface.

Currently the biological information included for access via the EHR prototype is gene and protein sequence variations information. There is no exact corresponding heading in the CCR specification which can be used explicitly to include these sequence variations. The mutation information could, however, fit in the 'Results' as one type of 'Laboratory results', which is a heading in CCR. Different mutation detection methods could also be fitted into 'Diagnostic procedures'.

Building a standards-based EHR system prototype to simulate the clinical system provides several advantages: First, it means that the underlying standard provides a base for many implementations. This avoids the need to examine each implementation separately, and makes the testing easier and the study more manageable. Secondly, any findings related to the standards-based EHR system prototype will contribute more to the domain. When the standard is modified to accommodate genetic information, existing implementations will have to be brought up to date if they are to comply with the standard, and new implementations will automatically conform. Furthermore, a standards-based EHR prototype provides potential for future communication, reuse and underpinning of a consistent data model.

The content and structure of OntoKBCF is accessible via the EHR interface and bio-health information can be personalised according to patients' parameters. Currently in this prototype patients' parameters include age, sex, ethnicity, and existing bio-health status, for example: patient's symptoms or diagnostic procedures. The structure of OntoKBCF can be used as a single explanation of semantic relationship for simple concept (for example hierarchy of coughing) and also can be used as a group explanation for a complex concept, for example: ***Patient_CF_with_Gly542X*** includes hierarchies of ***Patient_CF_with_Gly542X, Gly542X*** and ***Gly542***.

Explanation in OntoKBCF is more comprehensive compared with what is accessible via the EHR prototype interface: e.g. the affected gene or protein name, affected codon, mutation type. There are two reasons why not all of these explanations have been transferred through the EHR system prototype interface: 1) we have not transferred explanations that would either repeat existing ones or ones which are too similar; 2) property descriptions from the Protégé-OWL environment can currently only be used automatically to a very limited extent. Explanations that utilise properties can only be transferred manually.

6.2.4 Comparison of the EHR prototype with similar research

Hoffman[29] recommended incorporating clinically significant genomic information into the Electronic Medical Record (EMR). However, in that paper he did not describe how to fit specific genomic information into an EMR. He described more what was available and what the problems were but did not provide a concrete solution. The current research has not solved all the problems of bringing genetic information into a clinical setting, however. This research has, rather, explored how to organise biological information (mainly sequence variations) and health information in a formal, structured way and how to fit the content and structure in the knowledge base into an EHR prototype.

Hoffman, in another study[51], provided a clinical bioinformatics term/vocabulary support for Hospital Information System (HIS). In this research a bio-health knowledge base works as an active support for the EHR system prototype. In Hoffman's research rich terms are listed to describe findings by molecular diagnostic procedures and then imported into a commercial HIS system: his work can therefore be considered to be more horizontal; whereas OntoKBCF is on a vertical level, connecting the elementary concepts in biological and health fields, and it plays an active role after being embedded into an EHR system prototype.

Tange's[184] review of electronic medical record systems, was concerned with medical narrative data. This type of content has been involved in the research; however, it is not the focus. Most of the systems mentioned in the review used coded terms, controlled vocabularies or free text for medical narrative data; whereas both controlled vocabularies and structured concepts have been used in this research: these are in formal format to help in computer processing.

The Infobuttons[44, 49] research and this research share some characteristics, are reviewed in Chapter 2. My solution has been to construct a knowledge base prototype, from which knowledge facts have been organised formally and structurally and are personalised according to patients' parameters (such as age, sex, ethnicity and existing bio-health status) within an EHR system prototype. However, Infobuttons works by building an HTTP request according to specific clinical context and generating different URL hyperlinks for different knowledge sources. Their solution is more of a "portal" to relevant links rather than a final stop. The results of this research include hierarchies of concepts, and final candidate information related to a particular patient. Infobuttons researchers were trying to develop HL-7 based HTTP requests to tackle the communication problem[185]. Whereas one of the important characteristics of the solution in this research is the data sharing ability which was introduced through the semantic web technology. However, to update the knowledge base in this research is a challenge, and it would require sophisticated means to automatically keep the bio-information relevant. Infobuttons has the advantage of retrieving information from third party knowledge resources and relies upon their maintenance. Moreover, this research is a "lab product" compared with "Infobuttons", which has been integrated into real CIS.

6.2.5 Findings related to the connection between the knowledge base and the EHR

In this research the connection between the knowledge base prototype and the record prototype is also important, i.e., to make the formal and structured information accessible via the interface of the EHR system prototype is almost an application and research project in itself. Through the connection work it was realised that it is not currently realistic to connect the two parts by using simple query results, which would have made a clean interface between the two major components. Therefore an EHR related structure had to be constructed in OntoKBCF. This structure works as another axis in OntoKBCF with which to organise the major concepts. This structure is vital for establishing an automatic connection and making the hierarchy and content of OntoKBCF accessible through the EHR interface.

Through the EHR prototype interface, the hierarchy of basic concepts and combined concepts in OntoKBCF can be made accessible. This can be done both manually and automatically; however most of the knowledge facts in OntoKBCF at this time can only be used manually through the EHR. The reason for this is that the property descriptions and logic representations, whilst easily understood by human users, cannot be interpreted by the machine (i.e. the software) as it is not yet sufficiently developed to permit automatic processing.

6.3 Research significance

This research is a starting point for including bio-health information in the clinical system. The sequence variation data can be represented and organised and also can be used in this way. Explanation of mutation position, amino acid or nucleotide may be used to support non-specialist doctors in interpreting mutation in their consultation with patients. The degree to which this helps in clinical practice will need to be formally tested in future work.

This research is also a preliminary exploration of an active EHR system. To offer customised information automatically is one characteristic of what might be called an active EHR system, moving beyond a passive repository of clinical data. The biological and health information can be customised according to patient's parameters. It is consistent with the direction of EHR system's development: from management system to clinical system, from passive service to active service. The customisation of bio-health information is a potential help for doctor's information overload as it filters out what may be irrelevant.

This is also a starting point to connect a knowledge base with an EHR prototype. Although this research uses cystic fibrosis as an exemplar, the structure and organisation of the bio-health information and the way it is used in the EHR prototype may be generalised to other diseases, which have a clear cause of gene mutation and available associated health information. This is a good foundation for future automatic connection between complex knowledge bases and the EHR system.

This research explores the organisation of a range of knowledge: from nucleotide to organismal phenotype, from basic bio-health concepts to final knowledge facts in the same knowledge base. OntoKBCF plays an active (offer personalised information) and supporting (display hierarchies of concepts) role via the EHR prototype. The way bio-health information is organised in OntoKBCF and the way the content of the knowledge base is mapped to the EHR prototype are feasible.

Although there is a lack of solid foundation for plotting the precise pathway between biological base and pathologic status, even for cystic fibrosis, the two ends are loosely coupled in OntoKBCF. This research provides a foundation for creating a way to explore the relationships between biological information and disease status by bringing biological information into a clinical setting, which could be a platform for relationships study. This way is complementary to experimental validation, mathematics or statistics computing and modeling etc. and permits the same question to be seen from a different perspective.

6.4 Potential benefit

In this section the potential benefit for carrying out this research is explained. It should be noted that most of the benefit would be long term, which is not direct benefit from this research. Currently the research is a starting point.

To integrate biological information into clinical systems would benefit doctors, biomedical researchers, clinical system suppliers and standards development organisations in long term. Doctors would get more support in interpreting genetic tests results which in future might also influence treatment or prognosis guidance for patients, which depend on the domain research findings. Apart from that the prototype could be a tool for training and educating doctors who have to manage cystic fibrosis. For example, the tool may help to interpret the genetic test result correctly.

Biomedical researchers would get new perspectives in understanding relationships between biological information and disease status. The clinical system could be a platform for new precise evidences for connection research between biological information and disease status. For relationship research between biological information and disease status, this is also a complementary method.

Clinical system suppliers would benefit from the research for future systems integrating biological information. With the importance and huge amount of the biological information available revealed, biological information will inevitably be included in health records. This research is a preliminary exploration in this direction.

The research may benefit standards development organisations in integrating biological information into the new electronic record standard. Currently, at least for CCR, there is no explicit heading/structure for biological information. This research provides a starting point to do systematic research on how to integrate biological information into the new generation of record standards.

6.5 Research limitations

6.5.1 End user evaluation

The current research lacks formal end user evaluation of usability for the connection prototype. This will be required in any future work, but the focus of this early research has been on the feasibility of bringing the two types of information together and the functionality has been tested through the use of hypothetical records at this stage. Compared with the typical usability evaluation for clinical information systems[186], the future user usability evaluation for this research should focus on the content provided from OntoKBCF and the way the content is presented through the EHR interface, rather than being concerned with the efficiency of the EHR per se. Generally such evaluation will include answers to the following questions: 1) Is the content useful for a specific purpose? 2) Is the presentation method of the content, e.g. items in the OntoKBCF related boxes, tree view and extended tree view, helpful or are there better ways of presenting the same information? 3) How accurate is the presentation of the content and to what extent is information lost in different presentation styles? 4) How easy can the full content be displayed in front of a user and to what degree summarization and further explanations are required?

For question 3 for example, a major issue related with user's evaluation relates to the presentation method of the content from OntoKBCF which might not be sufficiently accurate or clear:

a) For example: CFTR AA2183_minus_G (a CFTR mutation) is a mutation which has relatively high prevalence in Italian CF patients. This connection has been used as a filter in the connection prototype (see Figure 5-10). If a patient has this mutation, then in the 'mutation related' box 'Italian' will be displayed. It suggests that the patient might (but not necessarily) be an Italian. However if a user is not familiar with the interface and/or the mechanism behind it, there is likely to be a misunderstanding. This might be corrected by training users to use the tool appropriately but changing the interface and the interaction might provide a clearer presentation of the original statement of knowledge to remove potential confusion;

b) Another example relates to the knowledge editor tool and to the extent it can represent uncertain knowledge.

The other major issue is essentially to do with the working context. In this research we have not yet tested the interaction between the display of the content from OntoKBCF and a user. The current preliminary method assumes a non-specialist user who would get the required support by providing knowledge hierarchies and showing location explanation for sequence variations; however the real users' requirements are likely to be much more complex and diverse depending on the context. The type of support required will vary from user to user. To test this would require the prototype to be much more comprehensive so as to permit meaningful usability evaluation. Full scale usability evaluation, looking at the cognitive and contextual issues, will be necessary to make a full implementation useful and usable for clinicians during the consultation.

6.5.2 Filters

The patient's demographic data was used as filters to customise the whole related set of information from OntoKBCF. At the moment the filters include age, sex, ethnicity and existing descriptions of individual patients' bio-health status. The filters rely on both the headings of the EHR system and the knowledge source of the exemplar disease. Age, sex and ethnicity are basic factors for medical records.

Originally some queries related to smoking (lifestyle) were tried by using Nucleotide, DDBJ, EMBL, OMIM and CFMDB, Genetic Home Reference, the Cystic Fibrosis Foundation[131], Mayo Clinic[134], and two text books[119, 120] and PubMed. The reason for making these queries was to see whether there was any biological information related to smoking in cystic fibrosis patients; if there was any result then the result could be included into OntoKBCF and smoking could become another filter. However, the lack of proof is symptomatic of the current gaps in domain knowledge, and as the knowledge matures then more filters can be introduced. The customised information has been proved by realistic but hypothetical test records; a more generalised conclusion will need an extensive evaluation.

6.5.3 Query of ontology

Given that OntoKBCF construction used a mature knowledge base development environment, it was presumed that it would be a straightforward step to query the content. Consequently, attempts were made to use Protégé-OWL 3.3.1, Sesame2.0[187] and TopBraid Composer 2.3.1[188] for this purpose, but, surprisingly, all failed to produce the required results or to deliver acceptable file export formats.

Because the direct query did not deliver the desired results, the idea of getting proper query results and to using the result in an EHR system was aborted. Ideal query results cannot be achieved by SPARQL either through Protégé-OWL or TopBraid Composer or SeRQL through Sesame, at the time of this research.

6.5.4 Information loss

Moving information between the loosely-coupled components of the overall system resulted in loss of information. From knowledge representation to knowledge presentation via the EHR interface some information loss occurs at each step. For example, in the knowledge representation part, imprecise content (such as "mostly" or "especially") cannot be represented precisely through Protégé-OWL.

Not all information within OntoKBCF could be mapped directly onto headings of CCR. For example: 'recurrent_coughing' is a subclass of 'coughing' in OntoKBCF domain part, both of which can be mapped to the CCR heading 'symptoms'. However, in the EHR part of OntoKBCF, 'recurrent_coughing' and 'coughing' cannot retain a hierarchical relationship if they are both under 'symptoms'. This conflicts with the domain part in OntoKBCF. The solution in this research is to keep the specific one, "recurrent_coughing', which is a pragmatic choice. However, this choice leads to information loss.

Not all the information in OntoKBCF can be transferred through the EHR interface without loss. For example, in OntoKBCF there is a final knowledge fact about the child cystic fibrosis patients: "has_manifestation some Heat_exhaustion". This means in child cystic fibrosis patients, patients will have heat exhaustion as a manifestation and may also have other manifestations. However, in the EHR system prototype interface 'Heat_exhaustion' will be displayed under the 'Age/Sex related' box as a possible candidate fact for child cystic fibrosis patients. This display does not provide a clear statement of the manifestation, i.e. an "all values from" (only have manifestation of heat exhaustion) or a "some values from" situation (may have other manifestations apart from heat exhaustion). It would be part of the user testing to determine the balance between full context being provided and minimising load through the EHR. In future, maybe, the presentation of candidate information from OntoKBCF through the EHR interface could include more detail, adding specific explanation or original description from the knowledge resources.

6.5.5 Knowledge sources

When OntoKBCF was constructed only sufficient knowledge sources were included to prove the concept rather than attempting to collect all comprehensive knowledge resources. Figure 6-1 provides a schematic overview of the layered knowledge base exploiting the metaphor of an EHR as a container and drawing from clinical and bio-information. The content produced for the knowledge base prototype, depicted by the horizontal levels in Figure 6-1, is sufficient to organise the relationships between the vertical levels for this research but not exhaustive. Other high quality knowledge sources which have not been included in current prototype are: for example, UpToDate, Clinical Evidence, Best Evidence and so on[152](p441).

There are many different approaches to achieving the research aim and even for the same approach there are still many different tools available. A pragmatic selection to accomplish the research aim was made; other possibilities are considered in the section on future research.

6.5.6 Biological research

Although significant progress has been made in explaining the biological essence of pathologic status, this is put into perspective when we consider the individual's lifestyle or the contextual environmental factors. The whole picture is then far from clear or complete even for monogenetic disease, not to mention polygenetic diseases. The science and technology have not yet reached the necessary level to explain all the mechanisms involved clearly, and it is doubtful they ever will. However, the recent discovery that diabetes may be due to a virus which attacks those with a genetic susceptibility only underlines the importance of understanding the relationships between the two types of knowledge. All of these factors make it far more difficult to decide what biological information should be included in clinical applications.

It is clear that although genetics has an important role to play, environmental and lifestyle factors are also very important given the complex, multi-facetted nature of health. Genetics, although fundamental, is also relatively stable and therefore long term: 'quick' and observable changes in conditions such as increases in diabetes or obesity will probably never be fully explained just by genetics.

Although not a limitation per se, one lesson learned was that it would appear prudent to do the connection work manually before considering automation. The manual process can give a clearer picture about the requirements for the connection and automation.

6.6 Future directions

In future user usability evaluation should include hierarchies and final knowledge facts evaluation by using credible numbers of users under a strict design. The tree view part provides a reference box for the user before the concept is added into a patient's EHR record. In the future full usability evaluation of this supportive role will be necessary in order to give feedback about the usability of knowledge structure and knowledge representation. Currently only the location information for mutation is included as a reference in the tree view box. In future it might be possible to add more types of information (apart from sequence location) so as to give the user a more comprehensive picture of the targeted item. Another related study would be to conduct a semantic check before an item is added from tree view to a patient's record in order to prevent mistakes.

A further task might include surveying the biological information needs of GPs, genetic specialists and other stakeholders, including patients. There is no direct biological information requirement from doctors available at the moment. There are many surveys of doctors' information needs, however, none have been found which focus on detailed classification of the contents that doctors need in their information seeking behaviour. All evidence is indirect [3, 18]

Currently the filters which were used in customising bio-health information in the EHR system prototype are static. Some of the filters will never change e.g., date of birth, sex, and ethnicity. Other characteristics such as blood group, iris and the immunisation status might be other types of filters that could be considered subject to availability. The systematic research of all the possible filters for genetic disease or other diseases might also be a future research topic.

Another future research direction related to CCR extension involves how different types of biological information could fit into the CCR structure; what new headings and new structures should be added; and what the detailed attributes and value ranges of these new pieces of information are.

Currently classes/concepts/hierarchies of the knowledge base are accessible through the EHR prototype. The parts related to property description and logic representation in ontology ("Asserted condition" part in Protégé-OWL interface) cannot be currently used directly, and this provides one area for further work.

There are two directions which can be explored on the technical level: how to use OntoKBCF output XML file and how to improve the ontology query results. In the current study the converted database table file from OntoKBCF was used for connection. There are many characteristics in the ontology which cannot be reached by present query, for example the properties description cannot be queried. This leaves space for future exploration.

Other directions need to be explored, for example the possibility of generalising to other genetic diseases by following the structure of OntoKBCF, the comparison with other EHR standards for the EHR-s prototype, and different knowledge structures or representation in OntoKBCF.

6.7 Summary

In this chapter the research results have been restated, interpreted and compared with other studies. Afterwards the research significance, limitations and future directions were discussed. This research has provided a useful starting point for including biological information directly into clinical applications. This is an explorative step towards an active EHR system; some of the present structure in OntoKBCF can be used automatically through the EHR system prototype, while certain parts can only be used manually. As a proof of concept the research achieved its aim. However, more research is required before the integration between the two types of knowledge becomes seamless and fully automated.

Chapter 7 Conclusions

7.1 Summary of the thesis

The aim of this research was to make formal and structured bio-health information accessible through a standards-based EHR system. A prototype was successfully constructed to prove the concept. Here I will give a brief review of what has been included in this thesis.

The first chapter gives the general context and background of this research and the motivations behind it: i.e., essentially the importance and rapid increase of biological information and the need to link it with health information in a formal and structured way such that it can be delivered in manageable quantities at the point of care. The increasing numbers of Electronic Health Record systems being used at the point of care suggested that this would be the tool used by doctors to filter the information and make it accessible. Furthermore, the recent emphasis on technical standards or specifications to underpin such systems suggested that the EHR and EHR-s should be standards-based. Although an area of increasing research, little could be found in the literature that combined these areas. The research aim and general solution follow from the motivation to explore the feasibility of accessing formal and structured biological information and associated health information by an ontology-based knowledge base prototype via a standards-based EHR system prototype. The major contributions of this research were presented.

The second chapter presents the literature review. Three primary parts include:

1) The rationale for this research and its position amongst existing similar work: e.g. because biological information is critical in understanding disease and the precise route between biological information and disease status is not completely clear; to bring biological information into simulated clinical setting provides a complementary method for relationship research between biological information and disease; there is no published research to the author's knowledge to bring ontology-based bio-health information, more precisely nucleotide sequence and amino acid sequence, to EHR systems;

2) The technical approach of this research: i.e. the reasons for selecting the technology used (i.e., ontology-based knowledge base and EHR system prototype) e.g., ontology makes the bio-health information reusable, sharable and extensible,

whereas EHR systems are one of most popular tools used in the clinical setting;

3) The choice of test domain, i.e., the candidate biological information resources, tools and exemplar disease-cystic fibrosis.

The third chapter is about research methodology. The research question, conceptual framework, research strategy, general procedures, research evaluation, requirements and functions of the knowledge base and the EHR prototype, reasons for choosing CCR as the standard and development environment introduction are included. This is an overview chapter about what was to be done. An ontology-based knowledge base prototype of cystic fibrosis was used to organise formal and structured bio-health information and a CCR-based EHR system prototype was built to interface with the knowledge base prototype.

The fourth chapter is one of two substantive content chapters, specifically about the construction of the ontology-based knowledge base prototype, named OntoKBCF. The detailed prototype construction work is included: i.e., OntoKBCF's scope, granularity, major axes, structures and specific construction procedures. Three examples are used to explain how knowledge facts have been organised in OntoKBCF. A summary of OntoKBCF, the rationale of the construction approach, design decisions and the limitations of OntoKBCF are also presented.

The fifth chapter is the other major substantive content chapter, specifically about the development of the EHR system prototype and its connection with OntoKBCF. Motivation for using the EHR system prototype is explained. Included here are the detailed procedures for building the EHR prototype; the connection with OntoKBCF; the mapping between CCR labels and OntoKBCF; and the degree to which the process can be automated. Following this, verification of the connection are shown through screenshots. OntoKBCF is shown to be actively accessible through the CCR-based EHR system prototype. The limitations of the EHR prototype and connection part are discussed.

The sixth chapter brings the parts together. The research results are shown and interpreted; research significance, implications, limitations and future directions are reviewed. As a proof of concept research study, the research aims have been achieved successfully in this study. It is nothing more, but nothing less than a good starting point to integrate formal and structured sequence variation information into the CCR-based

EHR system prototype and to present the customised bio-health information via the standards-based EHR prototype interface. The way to organise bio-health information in the ontology has been proved to be usable through the EHR prototype. The standards-based approach has the potential to provide a common underlying model for many EHR systems, irrespective of implementation. Basic concepts and combined concepts in OntoKBCF can be used automatically through the EHR prototype; the final knowledge facts in OntoKBCF can, to some extent, be used automatically. The knowledge base prototype can be connected with the standards-based EHR system prototype as perhaps one of many knowledge bases.

7.2 Restatement of aims

The general research goal of this research is to explore how to integrate different types of knowledge seamlessly into a clinical setting. The research aim is to explore the feasibility of offering formal, structured biological information and associated health information via a standards-based EHR system prototype.

7.3 Contributions of the research

The major contributions claimed for this research include:

- The bringing together of sequence variations information with health information using a standards-based EHR.
- The construction of an ontology-based cystic fibrosis bio-health model (the first of its kind) and knowledge base prototype.

7.4 Recommendation for future work

This research has raised the prospect of many future directions in need of further investigation. There are two directions in particular that I believe should be considered: 1) how to further personalise bio-health information through an EHR and 2) how to connect a stand-alone knowledge base into an EHR system.

7.4.1 Personalise bio-health information in the EHR

Personalised information is clearly beneficial for the patient, but it is also useful for doctors in relieving them from the burden of information overload caused by irrelevance. This research is a starting point in personalising bio-health information in the EHR prototype by filters. A method that might be used in future in personalising bio-health information through EHR would be to systematically investigate the suitability of other candidate filters, which may include headings from the EHR-s demographic data (apart from age, sex, ethnicity which are used in this research), epidemiologic data (such as life style and environmental factors), lab data (i.e. more biological genetic test results apart from just sequence variations) and clinical data (i.e. more types of diagnostic and therapeutic procedures etc.).

The filters are fundamental in providing personalised bio-health information. A method to carry out a systematic investigation of filters may include the following steps (using cystic fibrosis as an example): 1) to select all the factors closely related with cystic fibrosis that also can be found in an EHR, such as smoking, lung function test values, family history etc; 2) to consult with domain experts about what factors may correlate with a patient's molecular genetics change; 3) to retrieve literature to look for available evidence from both health and biological aspects according to suggestions from step 2; 4) review the retrieved literature and revise the filters to satisfy the available evidence. This method may also help to draw a clear domain map to indicate the current research gaps, to help experimental design and to choose domain research topics.

7.4.2 Stand-alone knowledge bases connection with EHR-s

My research has not done this. The constructed prototype has relied on two specifically custom built components being connected, albeit that the software tools (Protégé-OWL etc.) and specification (i.e. CCR) are independent artefacts. The same person has constructed both sides of the connection in this research. This would not happen in practice. Full semantic interoperability would be required to permit stand-alone knowledge bases to interchange effectively with other systems such as EHR-s, and this is still a challenge requiring formal agreements and standards as well as further research into semantic, structural and content representation.

The long term goal of harvesting appropriate knowledge and making it available at the point of care is a long journey. Although it will be important to make diverse knowledge bases plug-playable into an EHR system, there are still some critical barriers: 1) how to map and integrate knowledge in the knowledge bases with the many possible EHR structures automatically; 2) how to populate and present the knowledge through the

EHR interface without information loss. In this research semantic mapping has been achieved manually and in some extent automatically by using a database file format. Further investigation may consider XML format or message, for semantic interoperability between the knowledge bases and the EHR-s structure. Controlled vocabulary for both systems and knowledge bases, generic system structures, and detailed and scientific indexes for knowledge bases, are all important at the technique level. Meanwhile, to get clear and detailed user and information requirements will be equally important for connection. Figure 7-1 is a description of this recommendation.

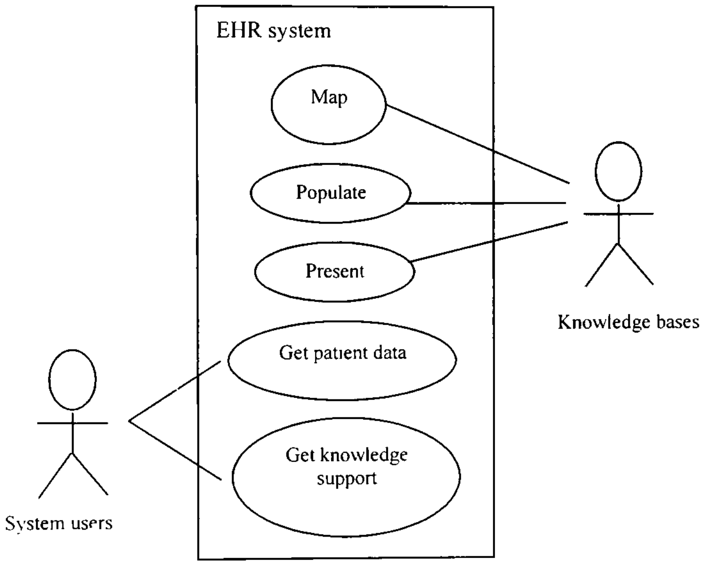


Figure 7- 1 Description of the EHR system with multi-knowledge bases

(Note: the knowledge base could be a cystic fibrosis knowledge base or a Huntington disease knowledge base etc; system user can be a doctor or other clinician)

7.5 Conclusions

The thesis has explored how to integrate formal and structured bio-health information into a simulated clinical setting. Sequence variations information and associated health information related to cystic fibrosis have been organised in an ontology-based knowledge base prototype- OntoKBCF. There are basic concepts, combined concepts and final knowledge facts in OntoKBCF. The hierarchy for basic and combined concepts in OntoKBCF can be made accessible manually and automatically through the CCR-based EHR system prototype. The representation of final knowledge facts (involving concept, property and description logic representation) can also be made accessible, although this is mostly manual with some automatic functionality. This work provides a way of delivering formal and structured bio-health information from

OntoKBCF via the CCR-based EHR system prototype.

This research has shown the feasibility of integrating sequence variation information into an EHR system prototype; usability tests in future research should concentrate on the relevance and most usable presentation of complex data. The integration of a wider range of biological information (currently the sequence variations knowledge supports include hierarchies for nucleotide mutation, amino acid change and corresponding nucleotide or amino acid position) will need further study. This is an exploration in integrating biological information into the EHR, but one has important potential in providing personalised health care service in future. In the long term view, such research should benefit doctors, biological researchers and patients.

The research suggests a need to expand the current CCR structure to include sequence variations. In CCR there is no specific label or example for biological information. The standards-based EHR prototype provides the common underlying data model for many implementations. This is also a starting point to present customised bio-health information via the CCR-based EHR. Sequence variations information has been included under the 'lab results' label in the EHR prototype. The procedures to connect a knowledge base with the EHR system prototype have been summarised. The bio-health information from OntoKBCF can be made accessible through the EHR system prototype interface and customised according to patient parameters. Filters include age, sex, ethnicity, existing bio-health description and diagnostic and treatment procedures. The customised information has been proved through the use of test records. It can be argued that to connect the knowledge base prototype with a standards-based EHR system prototype it is critical for further communication and standardisation. Much more work is required before plug-in stand-alone knowledge bases can be used within EHR-s without extensive manual intervention. This research should benefit standards development and EHR-s suppliers.

This work potentially lengthens the life of an ontology and has showed limitations in semantic web technology application and in previous ontology development: such as query limitation, limitation in using property description and logic symbols, semantic and logic conflicts between the domain knowledge and EHR structure etc. The way to organise the domain knowledge and EHR structure has been shown to work through the EHR system prototype. Both content and structure of the ontology are made accessible through the EHR prototype. Cystic fibrosis is used as an exemplar condition, however, the construction idea in ontology and connection method can be used for any other disease as long as there is a definite sequence variation and associated health information available. In this research, ontology has been used not only as a knowledge management tool, but also as an active knowledge resource in an EHR prototype. This research widens the semantic web application within the health care field.

In conclusion, this research has explored how to integrate formal and structured bio-health information into a simulated clinical setting, and in answering the research question it has raised a number of important others that need further research. The idea had certainly come of age and the direction of travel seems to be the right one. Even so, there is still much more research required before full implementations can deliver such information effectively at the point of care.

Appendix A: Introduction to RDF and OWL

RDF is a type of structure for representing information on the web[189]. It is metadata (i.e. data about data) used to represent machine processable information on the web[189] A recommended XML-based syntax from W3C RDF/XML- is used for exchange of information between applications (such as software agents)[189]. The structural unit for the RDF expression is an RDF triple: i.e., a subject, a predicate and an object, in that order[189]. Each RDF triple represents a statement between a subject and an object, connected by the predicate[189]. A set of RDF triples comprise an RDF graph, whose meaning contains the conjunction of all the triples' statement[189]. URI references and fragment identifiers are used in RDF naming[189].

The OWL ontology is an example of an RDF graph[92]. Although there are many forms in writing OWL ontology, it is the RDF graph that solely decides the meaning of an ontology[92]. This is that for the same OWL ontology, different syntactic forms share the same underlying set of RDF triples[92]

OWL Lite is an ontology language support only subset of OWL language constructs and specially designed to develop tool support for OWL, OWL DL and OWL Full support the same set of OWL language constructs, however the former has desirable functions for a reasoning system with more restrictive constraints[92] OWL Lite supports primarily a classification hierarchy and simpler constraints; OWL DL supports the maximum expressiveness and retains computational ability; OWL Full has maximum expressiveness and the syntactic freedom of RDF, with no computational guarantees[83].

Appendix B: Gene therapy of cystic fibrosis

The following text is from the publications$^{143, 146}$:

- Alton EWFW. Use of nonviral vectors for cystic fibrosis gene therapy. Proc Am Thorac Soc. 2004; 1:296-301.
- Flotte TR, Laube BL. Gene therapy in cystic fibrosis. CHEST. 2001; 120:124S-131S.

- Non_viral_vector_gene_therapy

["Annotation comments: This method appears more suitable for repeat dosing, but has been less effective. The transfer efficiency of non-viral vectors is currently low. Desirable properties of non-viral gene delivery systems, including biocompatibility (low toxicity and low immunogenicity), large adaptability with efficient nucleic delivery, ease of manufacturing (large quantities, high reproducibility, acceptable cost and simple storage conditions) and ease of clinical administration."]

  - Lipoplexes_vector_gene_therapy

["Annotation comments: Synthetic non-viral vectors can be formed by associating the nucleic acid sequences with cationic lipids to form lipoplexes. Cationic liposomes are lipid-DNA complexes, less efficient, do not stimulate inflammatory and immunologic responses."]

  - Lipopolyplexes_vector_gene_therapy

["Annotation comments: Synthetic non-viral vectors can be formed by associating the nucleic acid sequences with cationic lipids to form lipoplexes, with cationic polymers to form polyplexes, or with both cationic lipids and polymers to form lipopolyplexes."]

  - Polyplexes_vector_gene_therapy

["Annotation comments: Synthetic non-viral vectors can be formed by associating the nucleic acid sequences with cationic lipids to form lipoplexes, with cationic polymers to form polyplexes."]

- Viral_vector_gene_therapy

["Annotation comments: Unsuitable for repeat dosing, as immune response reduces the effectiveness of each subsequent dose. Fail to overcome host immune response to allow multiple readministration."]

  - Adeno_associated_virus_vector_gene_therapy

["Annotation comments: AAV vectors, they do not appear to induce inflammatory changes over a wide range of doses. The level of CFTR messenger RNA expression is difficult to ascertain with AAV vectors. AAV vectors appear to be safe and have superior duration profiles. AAV vectors establish latency in cells without inducing inflammation."]

  - Adenovirus_vector_gene_therapy

["Annotation comments: Ad vectors cause transient infection associated with acute inflammation."]

Appendix C: A brief summary of how cystic fibrosis may announce itself at different ages

The following text is from the book[119](p26):

➢ Harris A, Super M. Cystic fibrosis: the facts. 3rd ed. Oxford: Oxford University Press, 1995

- **In the new born:**

"Intestinal obstruction caused by meconium ileus or atresia; prolonged jaundice"

- **Infants:**

"Rectal prolapse; recurrent loose stools; distended abdomen; 'milky allergy'; recurrent chestiness, coughing, or wheezing; a salty taste to the sweat; poor weight gain, often associated with a ravenous appetite; unexplained dehydration

(Such symptoms are very common and may sometimes be mild, the diagnosis can be overlooked in infancy.)"

- **In older children:**

"After the diagnosis of CF in a brother or sister; incomplete intestinal obstruction; nasal polyps, especially if recurrent; bronchiectasis or recurrent chest infections; heat prostration; underweight child

(Occasionally the diagnosis may be missed for many years.)"

- **In adolescents or adults:**

"Delayed onset of puberty; sterile or azospermic males; infertile females with scanty cervical mucus; bronchiectasis"

Appendix D: Cochrane review conclusion about cystic fibrosis

The following text is from The Cochrane Library[129]:

- The Cocharane Collaboration. Cochrane review topics: cystic fibrosis. [cited 2006 July 7]; Available from: http://www.cochrane.org/reviews/en/topics/55.html

1. "*Bisphosphonates for osteoporosis in people with cystic fibrosis*: Intravenous pamidronate increases BMD at axial sites in people with CF, although it can cause severe bone pain in participants not receiving corticosteroids."
2. "*Chest physiotherapy compared to no chest physiotherapy for cystic fibrosis*: Airway clearance techniques have short-term effects in terms of increasing mucus transport."
3. "*Macrolide antibiotics for cystic fibrosis*: There is clear evidence from these studies of a small but significant improvement in respiratory function following treatment with azithromycin. The largest study employed a three times a week dose and, in this study, treatment with azithromycin was associated with a significant increase in mild adverse events."
4. "*Newborn screening for cystic fibrosis*: Nutritional benefits are apparent. Screening provides a potential opportunity for better pulmonary outcomes. Confounding factors such as severe genotype, pancreatic status and early acquisition with Pseudomonas aeruginosa have influenced long term pulmonary prognosis of people with CF in this study. Diagnosis through screening seems less expensive than a traditional diagnosis."
5. "*Non-invasive ventilation for cystic fibrosis*: Non-invasive ventilation may be a useful adjunct to other airway clearance techniques, particularly in people with CF who have difficulty expectorating sputum."
6. "*Prophylactic antibiotics for cystic fibrosis*: Anti-staphylococcal antibiotic prophylaxis leads to fewer children having isolates of Staphylococcus aureus, when commenced early in infancy and continued up to six years of age."
7. "*Recombinant human deoxyribonuclease for cystic fibrosis*: Therapy with rhDNase over a one-month period is associated with an improvement in lung function in CF. Results from a trial lasting six months also showed the same effect. Therapy over a two-year period (based on one trial) significantly improved FEV1 in children and there was a non-significant reduction in the risk of infective exacerbations. Voice alteration and rash appear to be the only adverse events reported with increased frequency in randomised controlled trials."

Appendix E: Most common CFTR mutations

The table is from the publication:

- Zielenski J, Tsui LC. Cystic fibrosis: genotypic and phenotypic variations. Anu Rev Genetics. 1995;29:777-807

Name of Mutation	Frequency	(%)	Population with the highest prevalence
[[Delta]]F508	28,948	-66	
G542X	1,062	-2.4	Spanish
G551D	717	-1.6	English
N1303K	589	-1.3	Italian
W1282X	536	-1.2	Jewish-Askhenazi
R553X	322	-0.7	German
621+1G->T	315	-0.7	French-Canadian
1717-1G->A	284	-0.6	Italian
R117H	133	-0.3	
R1162X	125	-0.3	Italian
R347P	106	-0.2	
3849+10kbC->T	104	-0.2	
[[Delta]]I507	93	-0.2	
394delTT	78	10-30%*	Nordic, Finnish
G85E	67		
R560T	67		
A455E	62		Dutch
1078delT	57		Celtic
2789+5G->A	54		Spanish
3659delC	54		
R334W	53		
1898+1G->T	53		
711+1G->T	49		French-Canadian
2183AA->G	40		Italian
3905insT	38	6-17%*	Swiss; Amish; Acadian
S549N	30		
2184delA	29		
Q359K/T360K		87.5%*	Jewish-Georgian
M1101K		69%*	Hutterite
Y122X		48%*	French, Reunion Island
1898+5G->T		30%	Chinese, Taiwan
3120+1G->A		11%	African-American
I148T		9.10%	French-Canadian

“The source of data is obtained from the CF Genetic Analysis Consortium (1994). The frequency is based on the screening of 43,849 CF chromosomes, although not all of them have been tested for the indicated mutations. The mutations are found in patients of Caucasian origin, except indicated otherwise. The geographic location (or ethnic group) with the highest prevalence is indicated for some of the mutations. A rough relative frequency (expressed in %,*) is given for those mutations studied in relatively small-size samples or in the indicated populations only [162]”

Appendix F: Characteristics of the most common cystic fibrosis mutations

The table is from the book[120](p11):

- Orenstein DM, Rosenstein BJ, Stern RC. Cystic fibrosis: medical care. Philadelphia: Lippincott Williams & Wilkins, 2000.

Mutations	Geographic/ethnic incidence	Other characteristics
[[Delta]]F508	70-75% in North America	Pancreatic insufficiency
G542X[1]	3.4% world wide	Pancreatic insufficiency; more meconium ileus
G551D[1]	2.4%worldwide	Pancreatic insufficiency
W1282X[1]	50-60% in Ashkenazi Jews; 2.1% worldwide	Pancreatic insufficiency
R553X[1]	1.3% worldwide	Pancreatic insufficiency
621+1G->T[1]	1.3% worldwide	Pancreatic insufficiency
1717-1G->A[1]	1.3% worldwide	Pancreatic insufficiency
R117H[1]	0.8% worldwide	Pancreatic sufficiency[2]; slightly lower sweat chloride; older age at diagnosis
R347P[1]		Pancreatic sufficiency[2];
3849+10kbC->T[1]	1.4% worldwide; 4% in Isreal	Pancreatic sufficiency[2]; normal sweat chloride; most males not sterile; lung disease varies from mild to severe
A455E[1]	3-7% in Netherlands; 0-0.2% in North America	Pancreatic sufficiency[2]; mild lung disease
R334W[1]		Pancreatic sufficiency[2]; older age at diagnosis
3905insT[1]	2.1% worldwide	Pancreatic insufficiency
P574H[1]		Pancreatic sufficiency[2];
Y563N[1]		Pancreatic sufficiency[2];
"[1]: Compound heterozygotes, in most cases, meaning these patients had one copy of the mutation noted and one other CF mutation (usually [[Delta]]F508)		
[2]: Pancreatic sufficiency in most, but not all, cases."		

Appendix G: Nomenclature of nucleotides and amino acids in OntoKBCF

Nucleotides are units of nucleic acid[64]. According to their chemical structure, nucleotides can be broken down into nucleosides and phosphate[190, 191]. The nucleosides can be further broken down into nitrogenous bases and ribose (as in RNA) or deoxyribose (as in DNA)[190, 191]. It is the nitrogenous base (Adenine, Thymine, Cytosine, Guanine, or Uracil) that decides the nucleotide type. "The one letter abbreviation can be used for either the bases alone or for the nucleotides containing them"[64] (p185). In OntoKBCF the one letter abbreviation name is used for nucleotide, which helps to provide compatibility with the mutation name used in the knowledge resources. When mutation occurs in the DNA or RNA chain, this is in the nitrogenous base, also named the nucleobase [192]; however, nucleotide also changes.

The International Union of Pure and Applied Chemistry (IUPAC) permits the naming of amino acids according to both a three letter abbreviation (e.g. Gly for Glycine) and a single letter abbreviation (e.g. G for Glycine). Naming within the ontology follows the three letter abbreviation name in order to avoid any confusion with the nucleotide name, which is a one letter abbreviation. The sub-hierarchy of amino acids uses the three letter naming convention (e.g. **Gly**) - both the full name and the single letter abbreviation are maintained as annotations. The sub-hierarchy of amino acid changes uses the three letter abbreviation name (e.g. ***Gly551Asp***) as does the sub-hierarchy of amino acid change location (e.g. ***Gly542***). Some mutation names which are from knowledge resources (reviewed academic publications in this study), such as G542X, are kept in the annotation.

In OntoKBCF nucleotide uses a one letter abbreviation name. Table G-1 lists the meaning of the abbreviations used in OntoKBCF[193].

Table G- 1 Full names list for nucleotides used in OntoKBCF

DNA	A	dAMP	Deoxyadenosine monophosphate
	G	dGMP	Deoxyguanosine monophosphate
	C	dCMP	Deoxycytidine monophosphate
	T	dTMP	Thymidine monophosphate
RNA	a	AMP	Adenosine monophosphate
	g	GMP	Guanosine monophosphate
	c	CMP	Cytidine monophosphate
	u	UMP	Uridine monophosphate

According to the IUPAC recommendation, 'X' can refer to any amino acid. However, according to the recommendation of protein sequence variants, 'X' should be used to designate a translation termination codon[163]. In the prototype, I followed the later recommendation to use 'X' to represent stop codon. As all of the 22 amino acids were listed already, the specific name (rather than 'X') could be used if any of them were needed.

Appendix H: Construction example: CFTR Gly542X

This example is about the description of one of most common CFTR mutations-Gly542X. CFTR mutations are related to ethnicity, which is another important filter used in individualising information. This example concerns the CFTR mutation Gly542X, which is located in exon 11 of CFTR gene. It has a higher prevalence in Spanish cystic fibrosis patients. There are 24 exons in the CFTR gene; a definition of exon is given in Appendix J. Cystic fibrosis patients with Gly542X may have pancreatic insufficiency, and many of them may have meconium ileus.

H1.1 Analyzing and dissecting the knowledge fact into basic concepts

This example is concerned with cystic fibrosis patients with a specific mutation (Gly542X), together with a high level explanation of the mutation, its relationship to ethnic characteristics and clinical manifestations.

Basic concepts include:

- cystic fibrosis patient: patient with Gly542X;
- amino acid substitution: Gly542X;
- exon: human CFTR gene exon 11;
- gene: human CFTR gene;
- ethnic population group: Spaniards;
- disease or syndrome: pancreatic insufficiency;
- disease: meconium ileus

H1.2 Selecting concepts/terms from UMLS and GO

According to the concepts/terms' definitions, we selected the concepts/terms and their superclasses for this piece of knowledge fact (Table H-1 to H-4).

Table H- 1 Amino acid and nucleotide related concepts from UMLS

Physical_object
Substance
Chemical
Amino_acid_peptide_or_protein
Amino_acids
Glycine
Nucleic_acid_nucleoside_or_nucleotide
DNA
Exon

Table H- 2 Conceptual related concepts from UMLS

Conceptual_entity
Idea_or_concept
Spatial_concept
Group
Population_group
Ethnic_group
Spaniards

Table H- 3 Diseases related concepts from UMLS

Phenomenon_or_process
Pathologic_function
Disease_or_syndrome
Diseases
Digestive_system_diseases
Gastrointestinal_diseases
Intestinal_diseases
Ileus
Pancreatic_insufficiency

Table H- 4 CFTR gene related concepts from UMLS

Physical_object
Anatomical_structure
Gene_or_genome
Human_CFTR_gene

H1.3 Creating and organising concepts

There are fewer molecular genetic concepts in UMLS than clinical concepts. Also, because GO has a focus which is totally different from OntoKBCF's, very few reference terms could be retrieved from it. While several classes were created to represent missing knowledge facts, the original identities and alternative vocabularies from UMLS were kept when available.

‘***Patient_CF_with_amino_acid_change***’ was created as a subclass of ‘***Patient_CF***’ It is also an interim and abstract class to associate cystic fibrosis patients with different amino acid changes.

‘***Patient_CF_with_Gly542X***’ was a subclass of ‘***Patient_CF_with_amino_acid_change***’ as shown in the Table H-5. Other sibling classes of '***Patient_CF_with_Gly542X***' included '***Patient_CF_with_Delta_Phe508***', '***Patient_CF_with_Asn1303Lys***' etc.

Table H- 5 Hierarchy of ***Patient_CF_with_Gly542X*** in OntoKBCF

Physical_object
Human_being
Patient_CF
Patient_CF_with_amino_acid_change
Patient_CF_with_Gly542X

'***Patient_CF_with_Gly542X***' was defined as an intersection of (1) ***Patient_CF_with_amino_acid_change***; (2) *has_mutational_property* some ***Gly542X***.

'*has_mutational_property*' is a property used to describe nucleotide mutation or amino acid change. There are 4 sub-properties: '*has_transversion_property*', '*has_transition_property*', '*has_deletion_property*', '*has_insertion_property*'. The former two are used specially for nucleotide mutations and the latter two can be used for both nucleotide mutation and amino acid change. For amino acid change, there is another type of change - substitution. We created '*substitute_from*' and '*substitute_with*' as sub properties of '*mutate*' to describe amino acid substitution.

The hierarchy of the "***Gly542X***" has been shown in Table H-6.

Table H- 6 Hierarchy of ***Gly542X*** in OntoKBCF

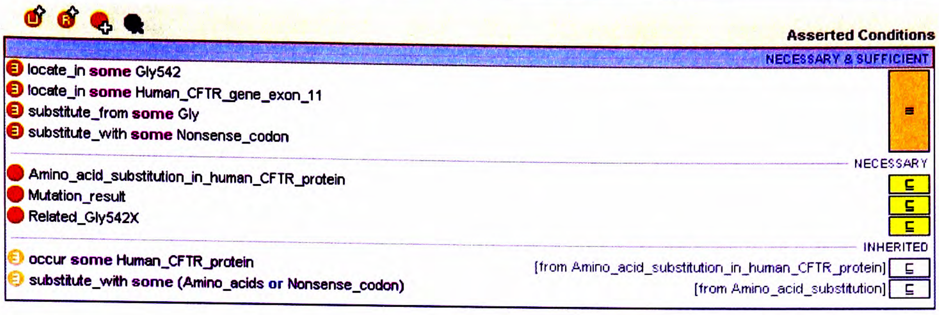
Phenomenon_or_process
Amino_acid_change
Amino_acid_substitution
Amino_acid_substitution_in_human_CFTR_protein
Gly542X

We defined '***Gly542X***' as an intersection of (1) *locate_in* some **Gly542**; (2) *locate_in* some ***Human_CFTR_gene_exon_11***; (3) *substitute_from* some **Glycine**; (4) *substitute_with* some **Nonsense_codon**. The screenshot for representation of '***Gly542X***' is in Figure H-1 with necessary and inherited conditions.

The '***Gly542***' was defined as a spatial concept first, and then it was used in definition of '***Gly542X***'.

Table H-7 shows hierarchy of '***Gly542***' in OntoKBCF.

'*locate_in*' is a property for location (number of nucleotide or amino acid).



Asserted Conditions

NECESSARY & SUFFICIENT

∃ locate_in **some** Gly542
∃ locate_in **some** Human_CFTR_gene_exon_11
∃ substitute_from **some** Gly
∃ substitute_with **some** Nonsense_codon ≡

NECESSARY

● Amino_acid_substitution_in_human_CFTR_protein ⊑
● Mutation_result ⊑
● Related_Gly542X ⊑

INHERITED

∃ occur **some** Human_CFTR_protein [from Amino_acid_substitution_in_human_CFTR_protein] ⊑
∃ substitute_with **some** (Amino_acids **or** Nonsense_codon) [from Amino_acid_substitution] ⊑

Figure H- 1 Representation of ***Gly542X***

The mutation '***Gly542X***' is located in exon 11 of CFTR gene. The exon 11 class hierarchy was created in OntoKBCF (Table H-8).

Table H- 7 Hierarchy of ***G542*** in OntoKBCF

Conceptual_entity
Idea_or_concept
Spatial_concept
Amino_acid_location_in_human_CFTR_protein
Gly542

Table H- 8 Hierarchy of ***Human_CFTR_gene_exon_11*** in OntoKBCF

Conceptual_entity
Idea_or_concept
Spatial_concept
Location_in_human_CFTR_gene
Human_CFTR_gene_exon
Human_CFTR_gene_exon_11

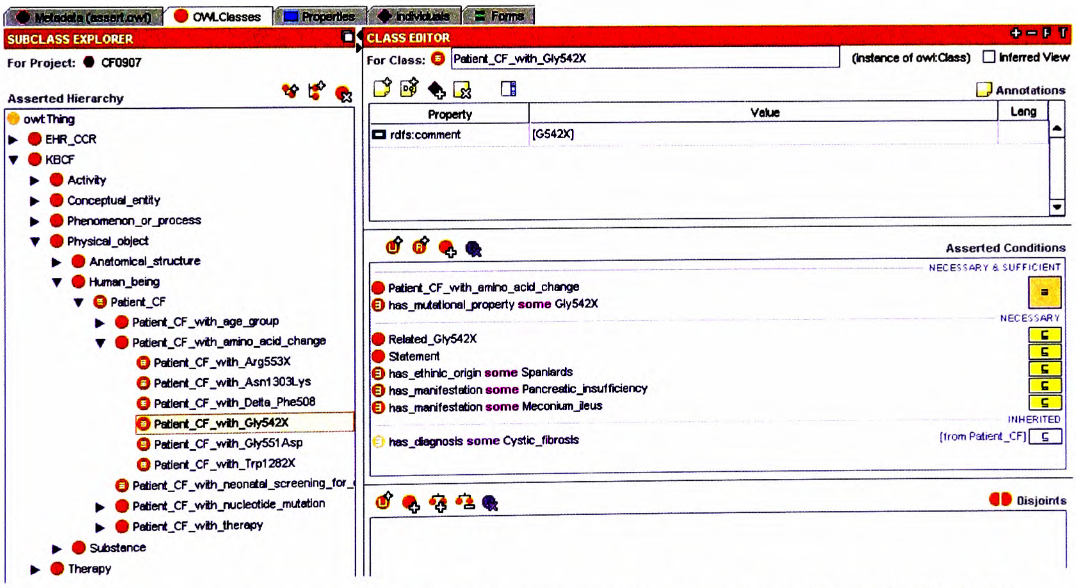
'**Nonsense_codon**' was used as one of subclasses of '**Nucleic_acid_nucleoside_or_nucleotide**' and with subclasses: '***DNA_nonsense_codon***' and '***mRNA_nonsense_codon***' (Table H-9). There is an equivalent class for '**Nonsense_codon**': '***X***'. Both of the classes are used to represent stop codon of amino acid in translation.

H1.4 Representing the knowledge fact

By combination of the classes and properties with logic symbols the final knowledge fact was represented under the '***Patient_CF_with_Gly542X***' with necessary conditions: (1) *has_ethinic_origin* some **Spaniards**; (2) *has_manifestation* some **Pancreatic_insufficiency**; (3) *has_manifestation* some **Meconium_ileus**. Figure H-2 shows the final representation and the hierarchical organisation of the ***Patient_CF_with_Gly542X***.

Table H- 9 Hierarchy and subclasses of **Nonsense_codon** in OntoKBCF

Substance
Chemical
Nucleic_acid_nucleoside_or_nucleotide
Nonsense_codon
DNA_nonsense_codon
TAA
TAG
TGA
mRNA_nonsense_codon
uaa
uag
uga
X



For Project: ● CF0907

Asserted Hierarchy

- owl:Thing
  - EHR_CCR
  - KBCF
    - Activity
    - Conceptual_entity
    - Phenomenon_or_process
    - Physical_object
      - Anatomical_structure
      - Human_being
        - Patient_CF
          - Patient_CF_with_age_group
          - Patient_CF_with_amino_acid_change
            - Patient_CF_with_Arg553X
            - Patient_CF_with_Asn1303Lys
            - Patient_CF_with_Delta_Phe508
            - Patient_CF_with_Gly542X
            - Patient_CF_with_Gly551Asp
            - Patient_CF_with_Trp1282X
          - Patient_CF_with_neonatal_screening_for_
          - Patient_CF_with_nucleotide_mutation
          - Patient_CF_with_therapy
      - Substance
    - Therapy

For Class: Patient_CF_with_Gly542X (instance of owl:Class) □ Inferred View

Property	Value	Lang
rdfs:comment	[G542X]	

NECESSARY & SUFFICIENT

Patient_CF_with_amino_acid_change
has_mutational_property **some** Gly542X

NECESSARY

Related_Gly542X
Statement
has_ethinic_origin **some** Spaniards
has_manifestation **some** Pancreatic_insufficiency
has_manifestation **some** Meconium_ileus

INHERITED

has_diagnosis **some** Cystic_fibrosis [from Patient_CF]

Figure H- 2 Whole screenshot for the representation of ***Patient_CF_with_Gly542X***

Appendix I: Construction example - Cochrane review topic description

Some Cochrane review conclusions about cystic fibrosis have been included in the knowledge base prototype. This is the part which can provide matching points with health information (such as symptoms). This health information has also been used in individualising patient information. The knowledge fact of Example 3 is: intravenous pamidronate treatment increases bone mineral density at axial sites in cystic fibrosis patient, although it can cause severe bone pain in participants not receiving corticosteroids.

I1.1 Analyzing and dissecting the knowledge fact into basic concepts

The positive effect is to increase bone density at axial sites for cystic fibrosis patients with intravenous pamidronate treatment; the side effect is severe bone pain for patients not using corticosteroids.

Basic concepts include:

- drug: pamidronate, corticosteroids;
- drug administration: intravenous drug administration;
- clinical characteristics: bone density;
- spatial: axial;
- symptom: bone pain;
- human being: cystic fibrosis patient;
- functional concept: increase.

I2.2 Selecting concepts from UMLS

According to their definitions the selected concepts and their superclasses from UMLS are listed in Table I-1 and I-2.

Table I- 1 Pharmacologic related concepts from UMLS

Physical_object
Substance
Pharmacologic_substance
Diphosphonates
Pamidronate
Adrenal_cortex_hormones

Table I- 2 Conceptual related concepts in UMLS

Conceptual_entity
Clinical_attribute
Bone_density
Sign_and_symptom
Bone_pain
Idea_or_concept
Functional_concept
Increase
Spatial_concept
Axial

I2.3 Creating and organising concepts

We created the '***Patient_CF_with_therapy***' as a subclass of '***Patient_CF***' in order to organise records for cystic fibrosis patients undergoing different types of therapy. '***Patient_CF_with_pamidronate_intravenous***' is a subclass of '***Patient_CF_with_therapy***'. The hierarchy is shown in Table I-3.

'***Patient_CF_with_pamidronate_intravenous***' was defined as an intersection of (1) ***Patient_CF_with_therapy***; (2) *treat_by* some ***Pamidronate_intravenous***.

'*treat_by*' is a property to describe patients with different treatments.

The hierarchy for '***Pamidronate_intravenous***' will be shown in Table I-4.

'***Pamidronate_intravenous***' was defined as an intersection of (1) **Pharmacotherapy**; (2) *administrate_ intravenously* some **Pamidronate** (Figure I-1).

Table I- 3 Hierarchy of ***Patient_CF_with_pamidronate_intravenous*** in OntoKBCF

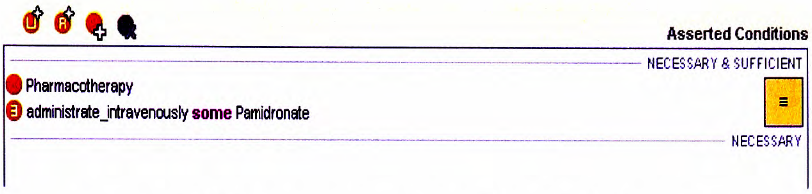
Physical_object
Human_being
Patient_CF
Patient_CF_with_therapy
Patient_CF_with_pamidronate_intravenous

'*administrate_intravenously*' is a sub property of '*administrate_drug_routes*' to describe intravenous drug administration, one of the drug administration routes.

'***Bone_density_increased_axially***' is a subclass of '***Bone_density_increased***', which is in turn a subclass of '**Bone_density**'. The hierarchy is shown in Table I-5.

Table I- 4 Hierarchy of ***Pamidronate_intravenous*** in OntoKBCF

Therapy
Pharmacotherapy
Pamidronate_intravenous
Pamidronate_intravenous_for_osteoporosis



Pharmacotherapy

administrate_intravenously **some** Pamidronate

Figure I- 1 Representation of ***Pamidronate_intravenous***

Table I- 5 Hierarchy of ***Bone_density_increased_axially*** in OntoKBCF

Conceptual_entity
Clinical_attribute
Bone_density
Bone_density_increased
Bone_density_increased_axially

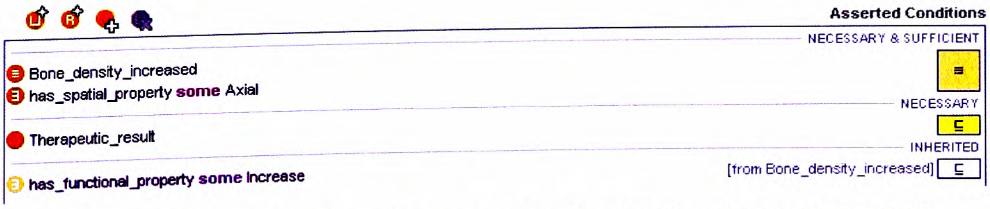
'***Bone_density_increased***' was defined as an intersection of (1) **Bone_density**; (2) *has_functional_property* some **increase**.

'*has_functional_property*' is a property used for objects with some functional concepts as property, such as decrease, insufficient etc.

'***Bone_density_increased_axially***' was defined as an intersection of (1) ***Bone_density_increased***; (2) *has_spatial_property* some **Axial** (Figure I-2).

'*has_spatial_property*' is a property used for some spatial property representation.

'***Bone_pain_severe***' was created as a subclass of '**Bone_pain**' and it was defined as an intersection of (1) **Bone_pain**; (2) *has_qualitative_property* some **Severe**. Table I-6 shows the hierarchy for "***Bone_pain_severe***".



Bone_density_increased

has_spatial_property **some** Axial

Therapeutic_result

has_functional_property **some** Increase

Figure I- 2 Representation of ***Bone_density_increased_axially*** in OntoKBCF '*has_qualitative_property*' is a property used for some qualitative properties.

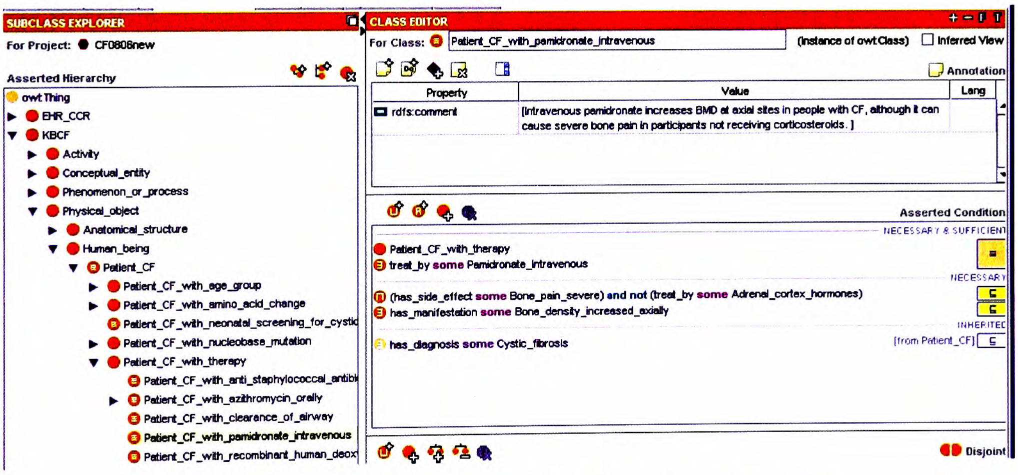
Table I- 6 Hierarchy of ***Bone_pain_severe*** in OntoKBCF

Conceptual_entity
Sign_and_symptom
Bone_pain
Bone_pain_severe

I2.4 Representing the knowledge fact

By combining the classes and properties with logic symbols the final knowledge fact was represented under '***Patient_CF_with_pamidronate_intravenous***' with the necessary conditions: (1) *(has_side_effect* <u>some</u> ***Bone_pain_severe)*** <u>and not</u> *(treat_by* <u>some</u> **Adrenal_cortex_hormones**); (2) *has_manifestation* <u>some</u> ***Bone_density_increased_axially***. Figure I-3 shows the representation of ***Patient_CF_with_pamidronate_intravenous***, including hierarchy and properties descriptions.

'*has_side_effect*' is a property used for patients or therapies with side effects.



For Project: ● CF0806new

```
owl:Thing
▶ EHR_CCR
▼ KBCF
    ▶ Activity
    ▶ Conceptual_entity
    ▶ Phenomenon_or_process
    ▼ Physical_object
        ▶ Anatomical_structure
        ▼ Human_being
            ▼ Patient_CF
                ▶ Patient_CF_with_age_group
                ▶ Patient_CF_with_amino_acid_change
                  Patient_CF_with_neonatal_screening_for_cystic
                ▶ Patient_CF_with_nucleobase_mutation
                ▼ Patient_CF_with_therapy
                      Patient_CF_with_anti_staphylococcal_antibi
                    ▶ Patient_CF_with_azithromycin_orally
                      Patient_CF_with_clearance_of_airway
                      Patient_CF_with_pamidronate_intravenous
                      Patient_CF_with_recombinant_human_deox
```

For Class: Patient_CF_with_pamidronate_intravenous (instance of owl:Class) □ Inferred View

Property	Value	Lang
rdfs:comment	[intravenous pamidronate increases BMD at axial sites in people with CF, although it can cause severe bone pain in participants not receiving corticosteroids.]	

NECESSARY & SUFFICIENT
Patient_CF_with_therapy
treat_by some Pamidronate_intravenous

NECESSARY
(has_side_effect some Bone_pain_severe) and not (treat_by some Adrenal_cortex_hormones)
has_manifestation some Bone_density_increased_axially

INHERITED
has_diagnosis some Cystic_fibrosis [from Patient_CF]

Figure I- 3 Whole screenshot for the representation of ***Patient_CF_with_pamidronate_intravenous***

Appendix J: Genetic vocabulary used in OntoKBCF

All these definition are from the publications[62, 64, 194, 195]:

- Campbell NA, Reece JB, Taylor MR, Simon EJ. Biology-concepts & connections. 5th ed. San Francisco, CA, USA: Benjamin Cummings, 2006.
- Skirton H, Patch C. Genetics for healthcare professionals-a lifestage approach. Oxford, UK: BIOS Scientific Publishers Limited, 2002.
- Raven PH, Johnson GB. Biology. 6th ed. New York, USA: McGraw-Hill Companies, 2002.
- Lewis R. Human genetics: concepts and applications. 2nd ed. Dubuque: Wm. C. Brown Publishers, 1997.

Chromosome	“The physical structures into which the DNA is packaged within the nucleus of cells. There are 46 chromosomes, 23 pairs in human somatic cell. Pairs numbered 1 to 22 are identical between males and females. There are two types of sex chromosome: X chromosome (females have two and males have one) and Y chromosome (only males have one).”
Nucleic acid	“A polymer consists of many nucleotide monomers; works as a blueprint and instruction for proteins, and for all cellular structures and functions through the actions of proteins. There are two types of nucleic acids, DNA (deoxyribonucleic acid) and RNA (ribonucleic acid).”
DNA	“The genetic material that organisms can inherit from their parents; DNA sequence can decide the sequence of amino acids; a double-helix macromolecule consisting of nucleotide monomers with deoxyribose sugar and the nitrogenous bases: adenine (A), cytosine(C), guanine (G), and thymine (T).”
RNA	““Bridge” substance between DNA and protein; RNA sequence is transcripted from DNA sequence and is translated into amino acids sequences; a type of nucleic acid consisting of nucleotide monomers with a ribose sugar and the nitrogenous bases adenine (A), cytosine(C), guanine (G), and uracil (U).”
Gene	“The fundamental physical and functional unit of heredity; a gene can be a sequence of DNA nucleotides in a chromosome; a gene can encode a protein, or tRNA, or rRNA molecule, or regulate the transcription. Genes can decide many traits of organisms.”
Codon	“The fundamental unit of the genetic code; a three-nucleotide sequence in mRNA that decides a particular amino acid or polypeptide termination signal; most of the amino acids are encoded by more than one codon.”
Mutation	“A permanent change in a cell's DNA normally involves change in biochemical composition; includes changes in nucleotide sequence, gene position, gene loss or duplication, and insertion of foreign sequences.”
Genotype	“An individual's genetic constitution underlying a single trait or set of traits.”
Phenotype	“The realised expression of the genotype; the physical appearance or functional expression of a trait.”
Transcription	“To synthesise RNA through reading DNA sequence, new formed RNA is complementary to the DNA template.”
Translation:	“To synthesise and assemble protein from codons in the RNA sequences. Genetic information transfers from nucleic acid to protein in this process.”
Exon	“A segment of gene sequence, which can code protein producing. Exons and introns are located alternatively in the gene. Exon also named as coding DNA.”

Intron	“A segment of gene sequence, which can't be translated into protein. Intron will be cut off during protein producing and only exon guides protein producing.”
Nonsense codon	“It also named as stop codon or termination codon. They are composed of three nucleotides in DNA or RNA. The codon will lead that synthesis of polypeptide chain stops prematurely.”

Appendix K: Events that occur when healthcare professional interacts with EHR and EHR system

The following description is from the book:

- Van de Velde R, Degoulet P. Clinical information systems: a component-based approach. New York: Springer-Verlag New York, 2003.

From Van de Velde and Gegoulet's[107](p94) these events occur when healthcare professionals interact with EHR:

> "(1) relate to a single patient; (2) are registered at one moment; (3) are discovered at one moment; (4) originate with and are under the responsibility, authorisation, and authentication of a responsible physician"

The following items need to be identified for each event[107] (p94):

> "(1) the subject of care; (2) the location of the patient; (3) the health service provider; (4) the time period; (5) the specialist service request or requester; (6) the source of the corresponding information; (7) the copy destination; (8) the item version; (9) the access and modification rights"

An EHR system has to offer following features[107](p118):

> "(1) a problem list; (2) the ability to measure the patient's health status and functional levels; (3) a tool to document clinical reasoning and rationale; (4) a longitudinal view that provides timely linkages with other patient records; (5) guaranteed confidentiality, privacy and audit trails; (6) continuous access for authorised users; (7) support for simultaneous multiple user views; (8) direct data entry by physicians; (9) support for practitioners in measuring or managing costs and improving quality; (10) flexibility in supporting existing or evolving needs of clinical specialties'

Appendix L: Major characteristics in the Protégé-OWL converted table

There were several major characteristics in the data fields of the Protégé-OWL 3.3.1 converted table. These data fields' characteristics were critical for the data to be extracted from the converted table and to be populated into the different tables used by EHR. For example:

1) for class in OntoKBCF the data fields would follow the form, "frame_type" = 9, "slot" = 2002, and "value_type" =3;

2) "value_type"—3 was for concepts and annotation, some internal strings might be represented with logic symbols;

3) "frame_type"—9 was for concepts and annotations, 13 was for property, 15, 16, 21, 22, 26, 30 were for internal strings, 25 was for instances, 28 was for some annotations, 29 was for name spaces;

4) "slot"—2002 was for concepts and 9093 was for annotations, 9097 was for name spaces;

5) Query statement parameters: frame=1000, frame_type=9, slot=12570, value_type=3.

The classes displayed as "text type" for the first time, and then they were represented by their "frame" number. For example, for class 'Atresia', the "frame" was 11991, after the first time with the "text" name-"Atresia", 11991 used for the class afterwards each time the class was referred to in the table.

Appendix M: Major coding design decisions in building the EHR system prototype

Implementation order	Sub name	Functions
1	Loadlist	Locate database's location; load patients' records list
2	Patientlist Selected	Load FormMaster; load patient' demographic data and bio-health data; load OntoKBCF related list boxes (clear, check presence, add)
3	FormMaster load	Load selected patient record with bio-health candidate data lists; show date;
4	Bio-HealthAdd	Add patient's bio-health data and update the backstage tables
5	Bio-HealthDelete	Delete bio-health data from patients' records/ backstage tables/ list boxes
6	CheckPresence	Check if the target item is already in patients' records, if yes ignore and if no, add the item to record
7	LstOntoSelected	Clear tree view, load tree view, show the selected concepts' hierarchy
8	LoadTree	Load OntoKBCF hierarchy for the target concepts
9	BtnAgeNew	Add selected item to corresponding patient's bio-health list, update backstage tables, patient's list boxes and reload OntoKBCF related list boxes
10	AddNewPatient	Add new patient record and then update the backstage table

Appendix N: Major coding design decisions in extracting tables automatically from the original Protégé-OWL converted table

Implementation order	Sub name	Functions
1	ClearOntology	Clean the existing ontology table
2	LoadOntology	Set target database; load tables; construct automatic database and update it
2.1	ScanDB	Extract all concepts in OntoKBCF
2.2	LoadCategory	Load different bio-health data tables according to EHR structure in OntoKBCF
2.3	GetParents	Get OntoKBCF hierarchies for the concepts
2.4	GetStatement	Get the concepts which need to be explained
2.5	GetRelationship	Get group concepts, whose hierarchies were used to explain concepts
2.6	ConstructDB	Construct the database according to loaded tables
2.7	UpdateOnto	Update OntoKBCF related tables- field "catid"

Reference

[1] Khoury M, Gwinn M, Yoon P, Dowling N, Moore C, Bradley L. The continuum of translation research in genomic medicine: how can we accelerate the appropriate integration of human genome discoveries into health care and disease prevention? Genet Med. 2007;9(10):665-674.

[2] Martin-Sanchez F, Iakovidis I, Nrager S, Maojo V, Groen Pd, Lei JVd, et al. Synergy between medical informatics and bioinformatics: facilitating genomic medicine for future health care. Journal of Biomedical Informatics. 2004;37(1):30-42.

[3] Collins FS, Green ED, Guttmacher AE, Mark S.Guyer on behalf of the US National Human Genome Research Institute. A vision for the future of genomics research. Nature. 2003;422:835-847.

[4] Drell D, Adamson A. Fast Forward to 2020: What to Expect in Molecular Medicine. [cited 2008 July 15th]; Available from: http://www.ornl.gov/sci/techresources/Human_Genome/medicine/tnty.shtml

[5] Galperin MY. The Molecular Biology Database Collection: 2005 update. Nucleic Acids Research. 2005;33:D5-24.

[6] Hall A, Walton G. Information overload within the health care system: a literature review. Health Info Libr J. 2004 2004 Jun;21(2):102-108.

[7] Davis DA, Ciurea I, Flanagan TM, Perrier L. Solving the information overload problem: a letter from Canada. MJA. 2004;180:S68-S71.

[8] Davies J, Fensel D, Harmelen FV. Towards the semantic web: Ontology-driven knowledge management. Chichester: John Wiley & Sons, 2003.

[9] Kalra D. Electronic health record standards. Methods Inf Med. 2006;45 Suppl 1:S136-S144.

[10] Semantic interoperability for better health and safer healthcare-Research and deployment road map for Europe. 2009 [cited 2009 July 3rd]; Available from: http://www.eurorec.org/files/filesPublic/2009semantic-health-report.pdf

[11] Stanford Center for Biomedical Informatics Research. Protégé-the Ontology Editor and Knowledge Acquisition System. [cited 2008 June 26th]; Available from: http://protege.stanford.edu/

[12] ASTM. ASTM E2369-05, Standard specification for continuity of care record(CCR); 2005.

[13] MedTerm Dictionary. [cited 2009 June 25th]; Available from: http://www.medterms.com/script/main/hp.asp

[14] Bochner BR. New technologies to assess genotype-phenotype relationships. Nat Rev Genet. 2003;4(4).309-314

[15] Miller PL. Opportunities at the intersection of bioinformatics and health informatics: a case study. J Am Med Inform Assoc. 2000:431-438.

[16] Blake J, Bult C. Beyond the data deluge: data integration and bio-ontologies. J Biomed Inform. 2006;39(3):314-320.

[17] Peltonen L, McKusick VA. Dissecting human disease in the postgenomic era. Science. 2001;291(5507):1224-1229.

[18] Shortliffe EH, Perreault LE, Wiederhold G, Fagan LM. Medical informatics : computer applications in health care and biomedicine. 2nd ed. New York: Springer, 2001.

[19] Hohenstein P, Fodde R. Of mice and (wo)men: genotype-phenotype correlations in BRCA1. Hum Mol Genet. 2003;12(2):R271-277.

[20] Jeanpierre C, Denamur E, Henry I, Cabanis MO, Luce S, Cecille A, et al. Identification of constitutional WT1 mutations, in patients with isolated diffuse mesangial sclerosis, and analysis of genotype/phenotype correlations by use of a computerized mutation database. Am J Hum Genet. 1998;62(4):824-833.

[21] Vermeire S, Wild G, Kocher K, Cousineau J, Dufresne L, Bitton A, et al. CARD15 Genetic Variation in a Quebec Population: Prevalence, Genotype-Phenotype Relationship, and Haplotype Structure. Am J Hum Genet. 2002; 71:74-83.

[22] Malin BA, Sweeney LA. Inferring genotype from clinical phenotype through a knowledge based algorithm. Pac Symp Biocomput. 2002:41-52.

[23] Cantor MN, Lussier YA. A knowledge framework for computational molecular-disease relationships in cancer. AMIA 2002; 2002, 101-105.

[24] Cantor MN, Lussier YA. Mining OMIM for Insight into Complex Diseases. Medinfo2004; 2004, 753-757.

[25] Tao Y, Friedman C, Lussier Y. Visualizing information across multidimensional post-genomic structured and textual databases. Bioinformatics. 2004;21(8):1659-1667.

[26] Lussier YA, Sarkar IN, Cantor M. An Integrative Model for In-Silico Clinical-Genomics Discovery Science. AMIA 2002; 2002, 469-473.

[27] Chen L, Friedman C. Extracting phenotypic information from the literature via natural language processing. In: Fieschi M, editor. MEDINFO 2004; 2004. IOS press, 758-762.

[28] Nguyen D. Partial least squares dimension reduction for microarray gene expression data with a censored response. Math Biosci. 2005;193(1):119-137.

[29] Hoffman MA. The genome-enabled electronic medical record. J Biomed Inform. 2007;40:44-46.

[30] Sax U, Schmidt S. Integration of genomic data in electronic health records: opportunities and dilemmas. Methods Inf Med. 2005;44:546-550.

[31] Robson B, Mushlin R. Genomic messaging system and DNA mark-up language for information-based personalized medicine with clinical and proteome research applications. J Proteome Res. 2004;3:930-948.

[32] HL7. V3 Messaging Standard. [cited 2009 12-16]; Available from: http://www.hl7.org/implement/standards/v3messages.cfm

[33] Bodenreider O, Stevens R. Bio-ontologies: current trends and future directions. Brief Bioinform. 2006;7(3):256-274.

[34] Burgun A. Desiderata for domain reference ontologies in biomedicine. J Biomed Inform. 2006;39(3):307-313.

[35] Rubin DL, Shah NH, Noy NF. Biomedical ontologies: a functional perspective. Brief Bioinform. 2007;9(1):75-90.

[36] Fredman D, Munns G, Rios D, Siegfried F, Siegfried M, Lenhard B, et al. HGVbase: a curated resource describing human DNA variation and phenotype relationships. Nucleic Acids Research,. 2004;32:D516-519.

[37] G.A.Thorisson, O.Lancaster, R.C.Free, R.K.Hastings, P.Sarmah, D.Dash, et al. HGVbaseG2P: a central genetic association database. Nucleic Acids Res. 2009;37:D797-802.

[38] Groth P, Pavlova N, Kalev I, Tonov S, Georgiev G, Pohlenz H, et al. PhenomicDB: a new cross-species genotype/phenotype resource. Nucleic Acids Res. 2007;35:D696-D699.

[39] Owen R, Altman R. Klein T. PharmGKB and the International Warfarin Pharmacogenetics Consortium: the changing role for pharmacogenomic databases and single-drug pharmacogenetics. Hum Mutat. 2008;29(4):456-460.

[40] Oliver DE, Rubin DL, Stuart JM, Hewett M, Klein TE, Altman RB. Ontology development for a pharmacogenetics knowledge base. Pac Symp Biocomput. 2002:65-76.

[41] Gupta A, Ludascher B, Grethe J, Martone M. Towards a formalization of disease-specific ontologies for neuroinformatics. Neural Networks. 2003;16(9):1277-1292.

[42] Mattes WB, Pettit SD, Sansone S-A, Bushel PR, Waters MD. Database Development in Toxicogenomics: Issues and Efforts. Environmental Health Perspectives. 2004;112(4):495-505.

[43] Wjst M, Immervoll T. An Internet linkage and mutation database for the complex phenotype asthma. Bioinformatics. 1998;14(9):827-828.

[44] Cimino JJ, Li JH, Allen M, Currie LM, Graham M. Janetzki V, et al. Practical consideration for exploiting the World Wide Web to create Infobuttons. MEDINFO2004; 2004. IOS Press, 277-281.

[45] Columbia University. The Infobutton Manager. [cited 2006 July 7]; Available from: http://www.dbmi.columbia.edu/cimino/Infobuttons.html

[46] Allen M, Currie LM, Bakken S, Patel VL, Cimino JJ. The classification of clinicians' information needs while using a clinical information system. In: Muusen M, editor. AMIA 2003; 2003, 26-30.

[47] Seol YH, Kaufman DR, Mendonca EA, Cimino JJ, Johnson SB. Scenario-based assessment of physicians' information needs. In: Fieschi M, editor. MEDINFO 2004; 2004. Amsterdam: IOS Press, 306-310.

[48] Seol YH, Johnson SB, Cimino JJ. Knowledge acquisition of generic queries for information retrieval. In: Cohane IS, editor. AMIA 2002; 2002, 1160.

[49] Cimino J. An integrated approach to computer-based decision support at the point of care. Trans Am Clin Climatol Assoc. 2007;118:273-288.

[50] Sheth A, Agrawal S, Lathem J, Oldham N, Wingate H, Yadav P, et al. Active semantic electronic medical record. In: Cruz I, Decker S, Allemang D, Preist C, Schwabe D, Mika P, et al., editors. 5th International Semantic Web Conference (ISWC 2006); 2006; Athens, GA, 913-926.

[51] Hoffman M, Arnoldi C, Chuang I. The clinical bioinformatics ontology: A curated semantic network utilizing RefSeq information. Pac Symp Biocomput. 2005;10:139-150.

[52] Kumar A, Yip L, Smith B, Grenon P. Bridging the Gap between Medical and Bioinformatics Using Formal Ontological Principles. Computers in Biology and Medicine. 2005:(In press).

[53] Seidl K. PheGe, the platform for exploring genotype-phenotype relations on cellular and organism level. Proc IEEE Comput Soc Bioinform Conf. 2002;1:79-86.

[54] Mitchell JA, Fun J, McCray AT. Design of Genetics Home Reference: a new NLM consumer heath resource. Journal of American Medical Informatics Association. 2004;11(6):439-447.

[55] NLM. Genetics Home Reference-your guide to understanding genetic conditions. [cited 2006 July 7]; Available from: http://ghr.nlm.nih.gov/

[56] Gene Ontology Consortium. Gene Ontology Home. [cited 2008 June 30th]; Available from: http://www.geneontology.org/

[57] NLM NIH. Unified Medical Language System. [cited 2008 June 30th]; Available from: http://www.nlm.nih.gov/research/umls/

[58] NIH NLM Programs and services 2000 (page50). 2000 [cited 2006 Mar. 9]; Available from: http://www.nlm.nih.gov/ocpl/anreports/fy2000.pdf

[59] NCBI NIH. GenBank. [cited 2008 May 29th]; Available from: http://www.ncbi.nlm.nih.gov/Genbank/

[60] PDB. Yearly growth of total structure in PDB. [cited 2008 May 29th]; Available from: http://www.rcsb.org/pdb/statistics/contentGrowthChart.do?content=total&seqid=100

[61] Baxevanis A. Using genomic databases for sequence-based biological discovery. Mol Med. 2003;9(9-12):185-192.

[62] Lewis R. Human genetics: concepts and applications. 2nd ed. Dubuque: Wm. C. Brown Publishers, 1997.

[63] Sadava D, Heller HC, Orians GH, Purves WK, Hillis DM. Life: The science of biology. 8th ed. Sunderland, MA, USA: Sinauer Associates, 2008.

[64] Campbell NA, Reece JB, Taylor MR, Simon EJ. Biology-concepts & connections. 5th ed. San Francisco, CA, USA: Benjamin Cummings, 2006.

[65] Guttmacher A, Collins F. Genomic medicine-A primer. N Engl J Med. 2002;347(19):1512-1520.

[66] U.S. Department of Energy Office of Science-Office of Biological and Environmental Research-Human Genome Program. Human Genome Project Information. [cited 2008 May 29th]; Available from: http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml

[67] Guttmacher AE, Collins FS. Genomics as a Probe for Disease Biology. N Engl J Med 2003;349:969-974.

[68] NCBI NIH. Nucleotide. [cited 2006 July 6]; Available from: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Nucleotide

[69] Research Organization of Information ans Systems-National Institute of Genetics. DNA Data Bank of Japan (DDBJ). [cited 2006 July 6]; Available from: http://www.ddbj.nig.ac.jp

[70] EBI. The EMBL Nucleotide Sequence Database. [cited 2006 July 6]; Available from: http://www.ebi.ac.uk/embl/

[71] Richards CS, Bale S, Bellissimo DB, Das S, Grody WW, Hegde MR, et al. ACMG recommendations for standards for interpretation and reporting of sequence variations: Revisions 2007. Genet Med. 2008;10(4):294-300.

[72] HGMD(The Human Gene Mutation Database Cardiff). Locus-Specific Mutation Databases. [cited 2009 Feb 10th]; Available from: http://www.hgmd.cf.ac.uk/docs/oth_mut.html

[73] Cotton RGH, Auerbach AD, Beckmann JS, Blumenfeld OO, Brookes AJ, Brown AF, et al. Recommendations for Locus-Specific Databases and Their Curation. Hum Mutat. 2007;29(1):2-5.

[74] Runz H, Dolle D, Schlitter A, Zschocke J. NPC-db, a Niemann-Pick type C disease gene variation database. Hum Mutat. 2008;29(3):345-350.

[75] Topel T, Hofestadt R, Scheible D, Trefz F. RAMEDIS: the rare metabolic diseases database. Appl Bioinformatics. 2006;5(2):115-118.

[76] Hise ME, Kattelmann K, M. P. Evidence-based clinical practice: dispelling the myths. Nutr Clin Pract. 2005 2005 Jun;20(3):294-302.

[77] Barnett GO, Barry MJ, Robb-Nicholson C, Morgan M. Overcoming information overload: an information system for the primary care physician. In: Fieschi M, Coiera E, Li Y-CJ, editors. MEDINFO2004; 2004; San Francisco,USA. IOS Press, 273-276.

[78] Hanka R, O'Brien C, Heathfield H, Buchan I. WAX ActiveLibrary: a tool to manage information overload. 32nd Hawaii International Conference on System Sciences; 1999, 4023-4031.

[79] Smith R. What clinical information do doctors need? BMJ. 1996;313:1062-1068.

[80] Mickan S, Askew D. What sort of evidence do we need in primary care? BMJ. 2006;332:619-620.

[81] Palmer SB. The Semantic Web: An Introduction. 2001 [cited 2006 Oct.22]; Available from: http://infomesh.net/2001/swintro/

[82] Updegrove A. The Semantic Web: an Interview with Tim Berners-Lee. 2005 [cited 2006 Oct. 22]; Available from: http://www.consortiuminfo.org/bulletins/semanticweb.php

[83] W3C. OWL Web Ontology Language Overview-W3C Recommendation 10 February 2004.DL McGuinness,F van Harmelen, editors [cited 2006 Oct. 10]; Available from:

http://www.w3.org/TR/owl-features/

[84] Hendler J, Lee TB, Miller E. Integrating Applications on the Semantic Web. Journal of the Institute of Electrical Engineers of Japan. 2002;122(10):676-680.

[85] Introduction. In: Davies J, Fensel D, Harmelen Fv, eds. *Towards the semantic web: Ontology -driven knowledge management*. West Sussex UK: John Wiley & Sons, 2003.

[86] Sowa JF. Building, sharing and merging ontologies. 2005 [cited 2006 May. 18]; Available from: http://www.jfsowa.com/ontology/ontoshar.htm

[87] Gruber T. What is an ontology? 1993 [cited 2006 Mar. 16]; Available from: http://www-ksl.stanford.edu/kst/what-is-an-ontology.html

[88] W3C. World Wide Web Consortium Issues RDF and OWL Recommendations-Semantic Web emerges as commercial-grade infrastructure for sharing data on the Web. [cited 2006 Oct. 10]; Available from: http://www.w3.org/2004/01/sws-pressrelease

[89] Wikipedia. Knowledge base. [cited 2008 June. 8]; Available from: http://en.wikipedia.org/wiki/Knowledge_base

[90] W3C. Semantic web. [cited 2006 Oct. 10]; Available from: http://www.w3.org/2001/sw/

[91] Swartz A. The Semantic Web In Breadth. 2002 [cited 2006 Oct. 22]; Available from: http://logicerror.com/semanticWeb-long

[92] Bechhofer S, van Harmelen F, Hendler J, Horrocks I, McGuinness D, Patel-Schneider P, et al. OWL Web Ontology Language Reference-W3C Recommendation 10 February 2004. [cited 2006 Oct. 10]; Available from: http://www.w3.org/TR/owl-ref/

[93] Wikipedia. Web ontology language. [cited 2006 Nov. 5th]; Available from: http://en.wikipedia.org/wiki/Web_Ontology_Language

[94] W3C. RDF/XML syntax specification (revised). [cited 2006 Nov. 5th]; Available from: http://www.w3.org/TR/rdf-syntax-grammar/

[95] Aranguren M, Bechhofer S, Lord P, Sattler U, Stevens R. Understanding and using the meaning of statements in a bio-ontology: recasting the Gene Ontology in OWL. BMC Bioinformatics. 2007;8:57.

[96] High-Tech dictionary. [cited 2006 Oct.11]; Available from: http://www.computeruser.com/resources/dictionary

[97] Lemaire F. Knowledge base, text and lexicon. In: Mars N, ed. *Towards very large knowledge bases: knowledge building & knowledge sharing 1995*. Oxford: IOS press, 1995;281-288.

[98] Bassiliades N, Vlahavas I. Knowledge-based system- techniques and applications. London. Academic Press, 2000.

[99] Benson T. Why general practitioners user computers and hospital doctors do not- Part 1: incentives. BMJ. 2002;325:1086-1089.

[100] Benson T. Why general practitioners use computers and hospital doctors do not Part 2:scalability. BMJ. 2002;325:1090-1093.

[101] Shortliffe EH, Cimino JJ. Biomedical informatics: Computer applications in health care and biomedicine. 3rd ed: Springer Science+Business Media, LLC, 2006.

[102] Committee on Imporving the Patient Record, Institute of Medicine. The computer-based patient record: an essential technology for health care: National Academy Press, 1997.

[103] Wu R, Straus S. Evidence for handheld electronic medical records in improving care: a systematic review. BMC Med Inform Decis Mak. 2006;6:26-30.

[104] NPfIT. Better information for health, where and when it's needed. [cited 2006 Oct.5]; Available from: http://www.connectingforhealth.nhs.uk

[105] Cross M. Keeping the NHS electronic spine on track. BMJ. 2006;332:656-658.

[106] Gunter TD, Terry NP. The emergence of national electronic health record architectures in the United States and Australia: Models, costs, and questions. J Med Internet Res. 2005 JAN-MAR 2005;7(1): art. no. e3.

[107] Van de Velde R, Degoulet P. Clinical information systems: a component-based approach. New York: Springer-Verlag New York, 2003.

[108] OpenEHR. OpenEHR. [cited 2008 May 5]; Available from: http://www.openehr.org/

[109] Children's Hospital Informatics Program, Dossia Consortium. IndivoHealth: Personally controlled health record. [cited 2008 June 9]; Available from: http://www.indivohealth.org/

[110] Román I. Toward the universal electronic healthcare record(EHR)- an ontology for the EHR. [cited 2006 July 6]; Available from: http://trajano.us.es/%7Eisabel/EHR/

[111]Records For Living Inc. HealthFrame- The Family Health Organizer. [cited 2008 June 12]; Available from: http://www.recordsforliving.com/HealthFrame/

[112] Chen R, Klein G. The openEHR Java reference implementation project. In: Kuhn Kea, editor. MEDINFO2007; 2007; Brisbane, Australia. IOS Press, 58-62.

[113] Ocean Informatics. [cited 2008 May 5th]; Available from:

http://oceaninformatics.biz/CMS/index.php

[114] Simons WW, Mandl KD, Kohane IS. The PING personally controlled electronic medical record system: technical architecture. J Am Med Inform Assoc. 2005;12(1):47-54.

[115] Hayrinen K, Saranto K, Nykanen P. Definition, structure, content, use and impacts of electronic health records: A reveiw of the research literature. Int J Med Inf. 2007.

[116] Ambre J, Guard R, Perveiler FM, Renner J, Rippen H. White paper: Criteria for assessing the quality of health information on the Internet; 1997.

[117] Walton G, Booth A. Exploiting knowledge in health services(p169). London: Facet Publishing, 2004.

[118] Rumsey S. How to find information- A guide for researchers (p168-169). England: Open University Press, 2004.

[119] Harris A, Super M. Cystic fibrosis: the facts. 3rd ed. Oxford: Oxford University Press, 1995.

[120] Orenstein DM, Rosenstein BJ, Stern RC. Cystic fibrosis: medical care. Philadelphia: Lippincott Williams & Wilkins, 2000.

[121] NLM NIH. PubMed. [cited 2006 July 6]; Available from: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed

[122] Kanz C, Aldebert P, Althorpe N, Baker W, Baldwin A, Bates K, et al. The EMBL Nucleotide Sequence Database. Nucleic Acids Research. 2005;33:D29-33.

[123] Tateno Y, Saitou N, Okubo K, Sugawara H, Gojobori T. DDBJ in collaboration with mass-sequencing teams on annotation. Nucleic Acids Research. 2005;33:D25-28.

[124] NCBI NIH. OMIM-Online Mendelian Inheritance in Man. [cited 2006 July 6]; Available from: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM

[125] Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase ofhumangenes and genetic disorders. Nucleic Acids Research. 2005;33:D514-517.

[126] Cystic Fibrosis Mutation Database. [cited 2008 May 29th]; Available from: http://www.genet.sickkids.on.ca/cftr

[127] NLM NIH. MeSH. [cited 2006 July 6]; Available from: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=mesh

[128] International statistical classification of diseases and related health problems. 10th ed. Geneva: WHO, 1992.

[129] The Cocharane Collaboration. Cochrane review topics: cystic fibrosis. [cited 2006 July 7]; Available from: http://www.cochrane.org/reviews/en/topics/55.html

[130] Jing X, Kay S, Hardiker NR. Bio-health information: a preliminary review of on-line cystic fibrosis resources. MIE2005; 2005, 21-26.

[131] Cystic Fibrosis Foundation. What is cystic fibrosis. [cited 2006 July 6]; Available from: http://www.cff.org/home

[132] Cystic Fibrosis Trust. What is CF. [cited 2006 July 6]; Available from: http://www.cftrust.org.uk/scope/page/view.go?layout=cftrust&pageid=28

[133] Merck & Co Inc. The Merck Manual of Diagnosis and Therapy, section 19, chapter 267, cystic fibrosis. [cited 2006 July 6]; Available from: http://www.merck.com/mrkshared/mmanual/section19/chapter267/267a.jsp

[134] Mayo Clinic Medical Services. Cystic Fibrosis. [cited 2006 July 6]; Available from: http://www.mayoclinic.com/health/cystic-fibrosis/DS00287/DSECTION=1

[135] U.S. Department of Energy (DOE) Human Genome Project Information. Genetic Disease Profile: Cystic Fibrosis. 2002 [cited 2006 July 6]; Available from: http://www.ornl.gov/sci/techresources/Human_Genome/posters/chromosome/cf.shtml

[136] Oxford University Gene Medicine UK CFGTC. Cystic fibrosis. [cited 2006 July 6]; Available from: http://users.ox.ac.uk/~genemed/

[137] Rowntree RK, Harris A. The phenotypic consequences of CFTR mutations. Annals of Human Genetics. 2003;67:471-485.

[138] Oxford University Gene Medicine UK CFGTC. Normal CFTR V Abnormal CFTR. [cited 2006 Oct.9]; Available from: http://users.ox.ac.uk/~genemed/

[139] Cuthbert AW. The cystic fibrosis gene. In: Shale D, ed. *Cystic fibrosis*. London: BMJ publishing group, 1996.

[140] Mueller RF, Young ID. Cystic fibrosis. In: Furn R, ed. *Emery's elements of medical genetics*. 11th ed. London: Churchill Livingstone, 2001;276-279.

[141] Kulczycki LL, Kostuch M, Bellanti JA. A clinical perspective of cystic fibrosis and new genetic findings: relationship of CFTR mutations to genotype-phenotype manifestations. Am J Med Genet A. 2003;116(3):262-267.

[142] Salvatore F, Scudiero O, Castaldo G. Genotype-phenotype correlation in cystic fibrosis: the role of modifier genes. Am J Med Genet. 2002;111(1):88-95.

[143] Flotte TR, Laube BL. Gene therapy in cystic fibrosis. CHEST. 2001;120:124S-131S.

[144] Montier T, Delepine P, Pichon C, Ferec C, Porteous DJ, Midoux P. Non-viral vectors in cystic fibrosis gene therapy: progress and challenges. TRENDS in Biotechnology. 2004;22(11):586-592.

[145] Lee TWR, Matthews DA, Blair GE. Novel molecular approaches to cystic fibrosis. Biochem J. 2005;387:1-15.

[146] Alton EWFW. Use of nonviral vectors for cystic fibrosis gene therapy. Proc Am Thorac Soc. 2004;1:296-301.

[147] UK Cystic Fibrosis Gene Therapy Consortium. Gene therapy. [cited 2009 July 5th]; Available from: http://www.cfgenetherapy.org.uk/gene_therapy/gene_therapy1.htm

[148] The Technology Source Archieves in the University of North Carolina. Proof of concept definition. [cited 2009 July 5th]; Available from: http://technologysource.org/extra/227/definition/5/

[149] Oates BJ. Researching information systems and computing. London: SAGE publications, 2006.

[150] Jing X, Kay S, Hardiker NR. Designing a Bio-health information assistant within an EHR. Healthcare Computing 2006; 2006; Harrogate England, 325-326.

[151] Jing X, Kay S, Hardiker N, Marley T. Ontology-based knowledge base model construction-OntoKBCF. MEDINFO 2007; 2007; Brisbane, Australia, 785-790.

[152] Ball MJ, Weaver CA, Kiel JM. Health information management systems: Cases, strategies, and solutions. 3rd ed. New York: Springer, 2004.

[153] DesRoches CM, Campbell EG, Rao SR, Donelan K, Ferris TG, Jha A, et al. Eletronic health records in ambulatory care - A national survey of physicians. N Engl J Med. 2008;359:50-60.

[154] Rector A. Foundations of the Semantic Web: Ontology Engineering. 2005 [cited 2006 Aug. 1st]; Available from: www.cs.man.ac.uk/~rector/modules/CS646/Ontology-building-2005-Lect-1.ppt

[155] Close HK. The information needs and behaviour of clinical researchers: a user-needs analysis. Health Information and Libraries Journal. 2005;22:96-106.

[156] Lappa E. Undertaking an information-needs analysis of the emergency-care physician to inform the role of the clinical librarian: a Greek perspective. Health Information and Libraries Journal. 2005;22:124-132.

[157] Casarett D, Karlawish J, Sankar P, Hirschman KB, Asch DA. Obtaining informed consent for clinical pain research: patients' concerns and information needs. Pain. 2001;92:71-79.

[158] Cimino JJ, Li JH, Bakken S, Patel VL. Theoretical, Empirical and Practical Approaches to Resolving the Unmet Information Needs of Clinical Information System Users. In: Cohane IS, editor. AMIA 2002; 2002, 170-174.

[159] Chen ES, Cimino JJ. Automated Discovery of Patient-Specific Clinician Information Needs Using Clinical Information System Log Files. In: Muusen M, editor. AMIA 2003; 2003, 145-149.

[160] Davies K. The information-seeking behaviour of doctors: a review of the evidence. Health Information and Libraries Journal. 2007;24:78-94.

[161] Arroll B, Pandit S, Kerins D, Tracey J, Kerse N. Use of information sources among New Zealand family physicians with high access to computers. J Fam Pract. 2002;51(8):706.

[162] Zielenski J, Tsui LC. Cystic fibrosis: genotypic and phenotypic variations. Anu Rev Genetics. 1995;29·777-807.

[163] den Dunnen JT. Recommendations for the description of sequence variants. [cited 2008 Sept 16th]; Available from: http://www.genomic.unimelb.edu.au/mdi/mutnomen/recs.html

[164] den Dunnen J. Nomenclature for the description of sequence variations. 2007 [cited 2008 Sept 16th]; Available from: http://www.hgvs.org/mutnomen/

[165] den Dunnen JT, Antonarakis SE. Recommendation for the description of protein sequence variants. Hum Mutat. 2000;15:7-12.

[166] Horridge M, Knublauch H, Rector A, Stevens R, Wroe C. A Practical Guide To Building OWL Ontologies Using The Prot´eg´e-OWL Plugin and CO-ODE Tools Edition 1.0. 2004 [cited 2006 Nov. 6th]; Available from: http://www.co-ode.org/resources/tutorials/ProtegeOWLTutorial.pdf

[167] Powell T, Srinivasan S, Nelson SJ, Hole WT, Roth L, Olenichev V. Tracking Meaning Over Time in the UMLS® Metathesaurus. Nov. 19th, 2002 [cited 2006 Nov. 2nd]; Available from: http://www.nlm.nih.gov/mesh/trackingmeaning.html

[168] Smith B. Home page of Barry Smith. [cited 2006 Mar. 8]; Available from: http://ontology.buffalo.edu/smith/

[169] ConnectingforHealth. GPSoC supplier systems. 2009 [cited 2009 June 29th]; Available from: http://www.connectingforhealth.nhs.uk/systemsandservices/gpsupport/gpsoc/systems/suppliers

[170] PRIMIS. System suppliers. 2008 [cited 2009 June 29th]; Available from: http://www.primis.nhs.uk/index.php/resources/useful-links/practice-system-suppliers

[171] Harriss C, Pringle M. Do general practice computer systems assist in medical audit? Family Practice. 1994;11(1):51-56.

[172] Beale T. openEHR a primer. 2007 [cited 2009 June 25th]; Available from: www.openehr.org/downloads/.../openEHR_primer_sep_2007.ppt

[173] O'Sullivan BP, Zwerdling RG, Dorkin HL, Comeau AM, Parad R. Early pulmonary manifestation of cystic fibrosis in children with deltaF508/R117H-7T genotype. Pediatrics. 2006;118:1260-1265.

[174] Chmiel JF, Drumm ML, Konstan MW, Ferkol TW, Kercsmar CM. Pitfall in the use of genotype analysis as the sole diagnosic criterion for cystic fibrosis. Pediatrics. 1999;103:823-826.

[175] Rosenbluth D, Goodenberger D. Cystic fibrosis in an elderly woman. Chest. 1997;112:1124-1126.

[176] Lohr C. Software testing. In: Thayer RH, Christensen MJ, eds. *Software engineering*. 3rd ed. Hoboken, New Jersey: IEEE Computer Society, 2005;413-420.

[177] Coward PD. A review of software testing. In: Thayer RH, Christensen MJ, eds. *Software engineering*. 3rd ed. Hoboken, New Jersey: IEEE Computer Society, 2005;421-430.

[178] Semantic reasoner. [cited 2009 11st Apr.]; Available from: http://en.wikipedia.org/wiki/Semantic_reasoner

[179] Racer. [cited 2009 11st Apr.]; Available from: http://www.racer-systems.com/

[180] Rubin D, Shah N, Noy N. Biomedical ontologies: a functional perspective. Brief Bioinform. 2008;9(1):75-90.

[181] Hernandez-Boussard T, Whirl-Carrillo M, Hebert JM, Gong L, Owen R, Gong M, et al. The pharmacogenetics and pharmacogenomics knowledge base: accentuating the knowledge. Nucleic Acids Research. 2008;36:D913-D918.

[182] About the UMLS Resources. 2004 [cited 2006 July 6]; Available from: http://www.nlm.nih.gov/research/umls/about_umls.html

[183] George RA, Smith TD, Callaghan S, Hardman L, Pierides C, Horaitis O, et al. General mutation databases: analysis and review. J Med Genet. 2008;45:65-70.

[184] Tange HJ, Hasman A, Robbe PFdV, Schouten HC. Medical narratives in electronic medical records. Int J Med Inf. 1997;46:7-29.

[185] Fiol GD, Rocha R. HL7 Infobutton standard API proposal. 2004 [cited 2006 Oct.16]; Available from: http://cslxinfmtcs.csmc.edu/hl7/arden/HL7-Infobutton-API-v2.2-20040224.doc

[186] Kushniruk AW, Patel VL. Cognitive and usability engineering methods for the evaluation of clinical information systems. J Biomed Inform. 2004;37(1):56-76.

[187] OpenRDF.org-home of Sesame. [cited 2008 Aug 6th]; Available from: http://www.openrdf.org/

[188] TopQuadrant-TopBraid Composer. [cited 2008 Aug 6th]; Available from: http://www.topquadrant.com/topbraid/composer/index.html

[189] W3C. Resource description framework (RDF): concepts and abstract syntax. [cited 2006 Nov. 5th]; Available from: http://www.w3.org/TR/rdf-concepts/

[190] Wikipedia. Nucleotide. [cited 2006 Nov. 6th]; Available from: http://en.wikipedia.org/wiki/Nucleotide

[191] Wikipedia. Nucleoside. [cited 2006 Nov. 6th]; Available from: http://en.wikipedia.org/wiki/Nucleoside

[192] Wikipedia. Nucleobase. [cited 2006 Nov. 6th]; Available from: http://en.wikipedia.org/wiki/Nucleobase

[193] Nucleobase. [cited 2009 July 5th]; Available from: http://en.wikipedia.org/wiki/Nucleobase

[194] Skirton H, Patch C. Genetics for healthcare professionals-a lifestage approach. Oxford, UK: BIOS Scientific Publishers Limited, 2002.

[195] Raven PH, Johnson GB. Biology. 6th ed. New York, USA: McGraw-Hill Companies, 2002.