




Universitat Autònoma de Barcelona

ADVERTIMENT. L'accés als continguts d'aquesta tesi queda condicionat a l'acceptació de les condicions d'ús establertes per la següent llicència Creative Commons:  http://cat.creativecommons.org/?page_id=184

ADVERTENCIA. El acceso a los contenidos de esta tesis queda condicionado a la aceptación de las condiciones de uso establecidas por la siguiente licencia Creative Commons:  <http://es.creativecommons.org/blog/licencias/>

WARNING. The access to the contents of this doctoral thesis it is limited to the acceptance of the use conditions set by the following Creative Commons license:  <https://creativecommons.org/licenses/?lang=en>

Departament Ciències de la Computació

Universitat Autònoma de Barcelona

Programa de doctorat en INFORMÀTICA

PhD Thesis

A Metamodel for Clinical Data Integration

Basis for a new EHR model driven by ontologies

Author: Raimundo Lozano Rubí

Supervisors: Prof. Mark Musen

Dr. Xavier Pastor Durán

Tutor: Dr. Josep Puyol-Gruart

July, 2016

Fdo. Prof. Mark Musen

Fdo. Dr. Xavier Pastor Durán

Fdo. Raimundo Lozano Rubí

A Marian.

Agradecimientos

Una tesis suele ser un viaje largo, y en mi caso lo ha sido especialmente. Lo que implica que también ha sido muy enriquecedor.

En primer lugar quiero dar las gracias a mis directores de tesis, Mark Musen y Xavier Pastor, por su guía a lo largo del camino. Mark Musen, con su trabajo en el dominio de las ontologías, ha sido una inspiración para mí desde el inicio. Xavier Pastor, con su conocimiento de la realidad de nuestras organizaciones sanitarias, ha conseguido que mantuviera los pies en el suelo y buscara soluciones a problemas reales, y ha sido un constante acicate para que llevara a término este proyecto. Pero Xavier ha sido mucho más que un director de tesis. Además de amigo, como jefe en el marco del Hospital Clínic de Barcelona durante más de 20 años, no sólo ha aceptado mis iniciativas y propuestas sino que las ha respaldado ante los demás.

También quiero dar las gracias a mi tutor, Josep Puyol-Gruart, al que le ha tocado bregar con los aspectos más administrativos y de procedimiento, siempre farragosos; y a Encarna Talavera, siempre dispuesta a aclarar mis dudas al respecto.

Durante estos años he tenido la oportunidad de colaborar con otros investigadores en varios proyectos relacionados con la tesis. Las conversaciones sobre las normas ISO 13606 y 13940 con Pablo Serrano y Adolfo Muñoz han sido especialmente clarificadoras e inspiradoras. Gran parte de la evaluación ha sido llevada a cabo implementando proyectos reales en el hospital, lo que no habría sido posible sin la colaboración de sus profesionales.

También distintos estudiantes y colaboradores han participado en estos proyectos y han ayudado en los sucesivos desarrollos, como Silvia Saiz, Felipe Geva y Pere Urbón. A ellos y a todos los que han circulado por nuestra unidad durante estos años mi agradecimiento, de todos he aprendido algo. Una mención especial se merecen nuestras dos secretarias durante estos años: Adelaida Farfán y Rosalía Aguayo. Siempre atentas conmigo y siempre dispuestas a echar una mano. Sin ellas el tiempo no habría cundido lo mismo.

Por último quiero dar las gracias a mi familia. A mis padres, por haberme iniciado en el camino. Siento profundamente que ninguno de los otros dos Raimundos de la familia pueda ver esta tesis acabada, sé que les hubiera hecho mucha ilusión. A mis hijos, mi mejor obra, por hacer que me sienta orgulloso cada día. Y de forma especial a Marian, quien ha sacrificado muchas aspiraciones personales para que yo pudiera alcanzar las mías. Gracias por tu apoyo y tu alegría constante.

Esta investigación ha sido parcialmente financiada por los proyectos PS09/02076 del PN 2008-2011 y PI12/01399, integrado en el Plan Estatal de Investigación Científica y Técnica y de Innovación 2013-2016 y cofinanciado por el ISCIII-Subdirección General de Evaluación y el Fondo Europeo de Desarrollo Regional (FEDER).

Abstract

The deployment of information systems in healthcare facilities has become widespread in recent decades and the main processes at Healthcare facilities are generally well supported. However, in spite of great advances in information and communication technologies domain during last years, current systems fail to provide true support to healthcare professionals in their daily practice and research activities. As a consequence of the variety of organizations providing healthcare and the heterogeneity of information systems used, current Electronic Health Record systems are not capable to show to healthcare professionals a conceptually consolidated view of the patients' health state. Patient's health data are fragmented inside information systems and over different information systems, and the professional should interpret and infer lacking relationships among them. In this scenario, semantic interoperability is pointed out by scientific community as an essential factor in achieving benefits from EHR systems to improve the quality and safety of patient care, public health, clinical research, and health service management.

In this thesis we propose OntoEHR, a conceptual architecture for a new semantically interoperable EHR system, focused on the clinical process and driven by ontologies. Conceptual and structural elements of the system are explicitly defined in OWL ontologies, conforming a declarative metamodel that drive all the system. Clinical data coming from different sources are stored and integrated in a clinical repository conforming to CEN/ISO 13606 standard, which is able to communicate clinical data using CEN/ISO 13606 extracts. Lastly, we propose a Problem Oriented Medical Record model, founded on

CEN/ISO 13940 standard, to represents patients' clinical data, assuring a safe and efficient continuity of care.

This thesis does not propose a specific and complete EHR system, but the foundation to build such systems.

Resumen

Durante las últimas décadas se ha extendido la implantación de sistemas de información en las organizaciones sanitarias, proporcionando un adecuado soporte a los principales procesos de las mismas. Sin embargo, a pesar de los avances producidos durante los últimos años en las tecnologías de la información y la comunicación, los sistemas actuales no son capaces de proporcionar un verdadero soporte a los profesionales sanitarios en su práctica diaria y sus actividades de investigación. Como consecuencia de la variedad de organizaciones sanitarias existentes y la heterogeneidad de los sistemas de información en uso, los sistemas actuales de Historia Clínica Electrónica no son capaces de mostrar a los profesionales sanitarios una visión conceptualmente consolidada del estado de salud de los pacientes. Los datos clínicos de los pacientes se encuentran fragmentados tanto entre diferentes sistemas de información como dentro de los mismos, de modo que los profesionales deben interpretar las relaciones entre los mismos así como inferir relaciones ausentes. En este escenario, la interoperabilidad semántica es considerada por la comunidad científica como un factor esencial para que los sistemas de HCE constituyan una ayuda para mejorar la calidad y seguridad de la atención a los pacientes, la salud pública, la investigación clínica y la gestión sanitaria.

En esta tesis proponemos OntoEHR, una arquitectura conceptual para un nuevo sistema de HCE semánticamente interoperable, enfocado sobre el proceso clínico y dirigido por ontologías. Tanto los elementos conceptuales como estructurales del sistema son definidos explícitamente mediante ontologías OWL, conforme a un metamodelo declarativo que dirige el sistema. Los datos clínicos procedentes de diferentes fuentes son almacenados e integrados en un repositorio clínico,

conforme con la norma CEN/ISO 13606, que es capaz de comunicar los datos clínicos mediante extractos CEN/ISO 13606. Por último, proponemos un modelo de Historia Clínica Orientada por Problemas, basada en la norma CEN/ISO 13940, para representar los datos clínicos de los pacientes, asegurando una continuidad asistencial segura y eficiente.

Esta tesis no propone ningún sistema de HCE específico y completo, sino las bases para construir tales sistemas.

Contents

1 Introduction	1
1.1 Processes in Healthcare	4
1.1.1 Healthcare and clinical process	4
1.1.2 Healthcare research process	7
1.1.3 Healthcare educational process	7
1.2 The Electronic Health Record	7
1.3 Semantic Interoperability	10
1.4 Ontologies	11
1.5 Terminologies.....	12
1.5.1 SNOMED CT.....	13
1.6 Data Structures	14
1.7 The CEN/ISO 13606 standard.....	15
1.8 The CEN/ISO 13940 standard.....	16
1.9 Structure of the Document	18
1.10 Derived Publications	18
1.11 Patents	22
2 State of the Art	25
2.1 EHR systems.....	25
2.2 Semantic Interoperability.....	27
2.3 Research support.....	29
2.4 Use of ontologies in EHR systems.....	31

2.5 Summary.....	33
3 Goals and Contributions	35
3.1 Objectives	35
3.2 Contributions	36
3.3 Hypothesis	37
3.3.1 General hypothesis	37
3.3.2 Specific hypothesis.....	37
3.4 Assumptions	38
3.5 Restrictions	38
3.6 Research Methodology	39
3.7 Summary.....	41
4 OntoEHR - The General Model.....	43
4.1 Functional requirements to cover	44
4.1.1 Essential capabilities.....	45
4.1.2 General capabilities	46
4.1.3 Research requirements.....	47
4.2 Workflow	48
4.3 OntoEHR Overview.....	48
4.4 System Architecture.....	51
4.5 Summary.....	54
5 OntoDDB.....	55
5.1 OWL-DB	56
5.2 OWL-DB Plug-in.....	61

5.3 OWL-DB OntoLoad.....	63
5.4 Summary.....	64
6 OntoCRF.....	65
6.1 Use case presentation.....	67
6.2 General architecture.....	70
6.3 Ontology authoring.....	71
6.4 OntoDDB-MM, The Metamodel.....	73
6.5 The Graphic User interface.....	78
6.6 Data extraction.....	79
6.7 Limitations.....	80
6.8 Summary.....	80
7 OntoCR.....	81
7.1 Meta-model description.....	84
7.2 Modeling data types: ISO 21090.....	85
7.3 Reference model: CEN/ISO 13606.....	86
7.4 Archetype model: CEN/ISO 13606.....	89
7.5 Linking the RM and Archetype model.....	90
7.6 Terminologies.....	91
7.7 Detailed archetype representation.....	92
7.8 Summary.....	93
8 POMR (A model of EHR).....	95
8.1 The Problem Oriented Medical Record.....	96
8.2 The Health Care Model Proposed.....	97

8.2.1 The concept of health problem	99
8.2.2 The work plan	103
8.2.3 The workflow	104
8.2.4 Actors and roles	106
8.2.5 Clinical work stations	108
8.3 Summary	109
9 Evaluation	111
9.1 OntoDDB	111
9.1.1 Repository evaluation	112
9.1.2 Semantic integration evaluation	114
9.2 OntoCRF	123
9.2.1 Real implementation evaluation	124
9.2.2 Performance evaluation	128
9.2.3 Usability evaluation	131
9.3 OntoCR	136
9.4 Summary	147
10 Conclusions and Future Work	149
10.1 Main Contributions	151
10.2 Hypotheses verification	154
10.3 Future Work	156
10.4 Conclusions	157
Bibliography	159
Web pages	172

List of Figures

4.1	General schema of OntoEHR	52
5.1	Basic components of OWL-DB	57
5.2	Types and properties in OWL-DB	58
5.3	Language representation in OWL-DB	59
5.4	Example of nested set model of trees	60
5.5	Class and Property hierarchies in OWL-DB.....	61
5.6	OWL-DB Plug-in.....	62
6.1	The list of clinical cases in CAPS.....	68
6.2	A clinical case in CAPS.....	69
6.3	General architecture of OntoCRF	71
6.4	Ontology edition with Protégé.....	72
6.5	OntoDDB-MM.....	74
6.6	Example of menu elements in OntoCRF	75
6.7	Example of form elements in OntoCRF	79
7.1	Overview of the general architecture of OntoCR	83
7.2	OWL representation of Reference Model.....	88

7.3	OWL representation of Archetype Model.....	89
7.4	OWL integration between Reference Model and Archetype model.....	90
7.5	OWL representation of vocabularies	92
8.1	The clinical process in continuity of care. From CEN/ISO 13940 standard.....	98
8.2	The care of patients' workflow. Modified from CEN/ISO 13940 standard	105
9.1	Searching content in SCOPE	117
9.2	Example of query using the ontology	120
9.3	Results obtained	122
9.4	Results of the computed SUS score, by each respondent. Results are displayed in ascending order. Dotted line marks an score of 68.....	133
9.5	Mindmap representing the tumour bank archetype.....	136
9.6	Representation of the clinical stage of a tumour in ADL format.....	138
9.7	Representation of the clinical stage of a tumour in OWL format	139
9.8	Example of application in OntoCR.....	140

List of Tables

3.1	Relation between the objectives and their corresponding contributions, hypotheses, assumptions, and restrictions	41
4.1	Different kinds of resources and its intended use	50
9.1	Results of OntoDDB performance evaluation	113
9.2	OntoCRF performance evaluation: Time required showing forms	130
9.3	Complains received by helpdesk support	134

Chapter 1

Introduction

Health Sciences in general, and Medicine in particular, are sciences based upon information and communication. Clinical practice and research processes consist mostly in collecting data, summarizing them and using information. This information, properly integrated with clinical knowledge, constitutes the base for decision support to take actions and generate new knowledge.

The deployment of information systems in healthcare facilities has become widespread in recent decades [26, 38]. Nowadays, it is a common business infrastructure in hospitals, medical offices and diagnosis centres. The main processes at Healthcare facilities are generally well supported, in particular in highly specialized hospitals, where high complexity healthcare delivery and major investment in high tech equipment for diagnosis and therapies requires ubiquitous access, immediacy, concurrency, security and continuous operation (24x7) of all information systems (IS).

However, in spite of great advances in the field of information and communication technologies (ICT) during last years, current systems fail to provide true support to healthcare professionals in their daily practice [8, 18, 45]. New features are often requested, including more registration of computable clinical data, “smart” gathering of patient data before clinical contact, intelligent

advice for the professional and the patient, offering relevance, pertinence, adequacy, appropriateness and a contextual user interface with interactive capabilities ready to support the clinical decision and the therapeutic intervention in real time at the point of care [59, 87]. Although there is a broad commercial offer of clinical information systems to support the patient management and the electronic patient record, they are mainly Enterprise Resource Planning systems (ERP). ERP systems are focused on the economic and administrative processes, and lack the needed functionality to manage clinical data.

Reusing existing clinical data for research purposes is not easy. Existing central data warehouses usually fail to support the creation of structured variables for research use [37], so it is necessary to build dedicated systems [27]. As a result, there is little institutional support for the collection of clinical data, especially for research, in health organizations.

The implementation of data repositories for research purposes has been reported to increase the capacity of a research team [37]. Some surveys show that individual organizations are progressing to the development, management, and use of clinical repositories as a means to support a broad array of research [54]. Although most researchers already use some software system to manage their data in electronic format, there continues to be widespread use of basic and general-purpose applications, such as spreadsheets, and additional support has become necessary for managing datasets. Interestingly, the barriers to acquiring currently available tools are most commonly related to financial burdens [3].

The situation in the Hospital Clínic of Barcelona (HCB) is one of these cases. The HCB has a long tradition in Biomedical research and stands as a benchmark institution both nationally and internationally [5, 75]. A research project cannot be longer understood without the ICT support in some extend. Nevertheless, the spreadsheet remains still as the “key tool” for research data management, as

financial limitations and lack of informatic expertise avoid acquiring of more complex tools. Continuous change is a characteristic of biomedicine domain, and building applications that can handle it is very expensive.

Other important aspect to consider is a growing tendency to provide health care through different healthcare facilities which have to collaborate. In this scenario, semantic interoperability is an essential factor in achieving benefits from Electronic Health Record systems (EHR) to improve the quality and safety of patient care, public health, clinical research, and health service management [92]. Communication between different facilities must be precise and lacking of whatever type of noise. In this context, standards are essential for the development and deployment of interoperable eHealth systems. Generic information models, clinical models, ontologies and terminologies have been identified as required artefacts to achieve semantic interoperability [57], but closer integration between these elements is needed [36, 92].

In this thesis we propose OntoEHR, a conceptual architecture for a new semantically interoperable EHR system, focused on the clinical process and driven by ontologies. Conceptual and structural elements of the system are explicitly defined in OWL [70] ontologies, conforming a declarative metamodel that drives all the system, from how to structure data to the GUI. The ontological nature of the system allows its seamless integration with existing knowledge, which becomes one more component. Clinical data coming from different sources are stored and integrated in a clinical repository conforming to CEN/ISO 13606 standard [EN 13606]. This repository is able to automatically incorporate external CEN/ISO 13606 extracts, and to generate CEN/ISO 13606 extracts with the data it contains. This thesis do not propose a concrete and complete EHR system, but the foundation to build such systems.

1.1 Processes in Healthcare

The main operations in all healthcare organizations are related to the interaction between subjects of care and healthcare professionals.

The generic definition of a process, outlined in ISO 9000:2005 [ISO 9000], is a "set of interrelated or interacting activities which transforms inputs into outputs". Processes are built up by activities that influence process objects, representing the inputs that are then, as process objects, transformed into outputs. Processes can be aggregated and/or subdivided into different parts that can be considered as processes by themselves. Three main types of processes in healthcare organizations have been identified [ISO 13940]:

- Healthcare and clinical processes
- Healthcare research processes
- Healthcare educational processes

1.1.1 Healthcare and clinical process

A healthcare process includes any set of activities related to the interaction between a subject of care and healthcare professionals. Inputs and outputs are health states represented by health issues, described as observed aspects of the health state.

The main “products” in healthcare are healthcare services (the end results of clinical processes). The subject of care is the receiver of these healthcare services as they improve or maintain the health state of that subject of care.

The clinical process is envisaged as consisting of four separate processes that can be executed separately, but that share information [ISO 12967-1].

1.1.1.1 Diagnostic consideration

Diagnostic consideration is a process in which facts about the patient are collected and analyzed in order to understand the health problems presented, the health state and the health needs.

This process implies that the clinician formulates the problems that are central to the patient establishing a “problem list”. This problem list will change through the clinical process as a consequence of the different healthcare activities. The problem list constitutes a basic tool for the clinician and should express a professional description of the patient condition.

The next step is to assess the needs for healthcare activities, aiming to identify health conditions and to treat already recognized health conditions.

1.1.1.2 Planning

Planning is a process during which activities to be conducted are planned along with the outcome expected and put into the care plan.

All active or latent health problems can demand the execution of different healthcare activities, e.g. healthcare investigations, drug prescription, surgical procedures, nurse activities, etc.

1.1.1.3 Executing

Executing is a process in which the plans are implemented and the actual interventions are conducted. Following the actual interventions conducted a set of results are obtained, which are seen as information about the patient condition.

1.1.1.4 Evaluation

Evaluation is a process in which you compare the expected outcome with the actual results, and draw conclusions (informed opinions) about the resultant conditions. The concept of 'clinical evaluation' includes both a comparison and an assessment. After completion of all activity elements in the care plan, the clinical process outcome (its effect on the health state) is observed, analyzed and described as one or more resultant conditions.

In the process of clinical evaluation can take part activities performed by healthcare professionals and activities automatically performed, e.g. by decision support processes.

1.1.1.5 Healthcare activities

The main healthcare activities used in clinical processes to transform the health state of a patient belong to one of the following categories:

Observe: To observe is to recognize a phenomenon, for example to observe an aspect of a health state as a health condition.

Assess: To assess is to form an opinion concerning the relevance of the observed conditions. Examples are to assess the cause and severity of an observed health condition and the assessment of the effect of treatment.

Plan: The care plan is the center around which the clinical process is delivered and covers all stages of the investigating and therapeutic activities' life cycles.

Take action: To take action is to perform/execute the investigations and treatments that are set out in the care plan.

1.1.2 Healthcare research process

The healthcare research process is a type of process in some healthcare organizations with an objective to contribute to clinical knowledge in general.

Related activities can be:

- Formulating conditions for cases selection.
- Selecting information elements.
- Data anonymization.
- Executing computation over data, for example statistical analysis.

Is an objective of this thesis to facilitate the use of clinical information for these purposes, letting aside the processes themselves.

1.1.3 Healthcare educational process

Another type of process in some healthcare organizations is the healthcare educational process with the aim to introduce and develop the knowledge and skills for healthcare. These processes can be targeted to healthcare professionals or to patients.

Is an objective of this thesis to facilitate the use of clinical information for these purposes, letting aside the processes themselves. The proposed model, based on ontologies, will facilitate the generation and dissemination of new knowledge.

1.2 The Electronic Health Record

Accordingly with the standard CEN/ISO 13940 [ISO 13940], a health record is a data repository regarding the health and healthcare of a subject of care. EHR systems currently in use merely records the resulting information from the interaction among a patient and healthcare professionals. The health record should carry out several functionalities. To act as a reminder and help in the

clinical management of the patient. To facilitate the communication among the different components of the clinical team and the continuity of care. To represents the care provided, enabling the retrospective analyses of clinical practice.

The essential function of the health record is to support the healthcare process, therefore should be considered the core of whatever clinical information system. Other secondary uses are teaching and research. Selected cases can be used for students learning and for professional training, and is habitual to use the health record data for research purposes. Finally, the health record has a legal consideration and can be used for auditing.

Paper-based records have been in existence for centuries and their gradual replacement by computer-based records has been slowly underway for over twenty years in western healthcare systems.

The US IOM report, Key Capabilities of an Electronic Health Record System [93], identified a set of 8 core care delivery functions that electronic health records systems should be capable of performing in order to promote greater safety, quality and efficiency in health care delivery:

- **Health information and data.** An EHR system must contain certain data about patients. Physicians and other care providers require having immediate access to key information - such as patients' diagnoses, allergies, lab test results, and medications.
- **Result management.** The ability for all providers participating in the care of a patient in multiple settings to quickly access new and past test results.
- **Order management.** The ability to enter and store orders for prescriptions, tests, and other.
- **Decision support.** Using reminders, prompts, and alerts, computerized decision-support systems would help improve compliance with best clinical practices, ensure regular screenings and

other preventive practices, identify possible drug interactions, and facilitate diagnoses and treatments.

- **Electronic communication and connectivity.** Efficient, secure, and readily accessible communication -- among health care team members and other care partners (e.g., laboratory, radiology, pharmacy) and with patients -- is critical to the provision of quality health care.
- **Patient support.** Tools that give patients access to their health records, provide interactive patient education, and help them carry out home-monitoring and self-testing.
- **Administrative processes.** Computerized administrative tools, such as scheduling systems for hospital admissions, inpatient and outpatient procedures, and visits.
- **Reporting and Population Health Management.** Electronic data storage that employs uniform data standards will enable health care organizations to respond more quickly to federal, state, and private reporting requirements, including those that support patient safety and disease surveillance.

The EHR provide new possibilities for secondary uses of the health data [79]:

- Clinical and translational research. Data recorded in health records is used for patient-oriented research, epidemiologic and behavioral studies and outcomes research and health services research.
- Public health, e.g. to prevent the spread of diseases, promote and encourage healthy behaviors and assure the quality and accessibility of health services.
- Quality measurement and improvement: improving experience of care, improving health of population and reducing costs [11]

Although the growing amount of data in EHR systems provides unprecedented opportunity for its re-use, there are many caveats to the use of such data. EHR data from clinical settings may be inaccurate, incomplete, transformed in ways that undermine their meaning, unrecoverable for research, of unknown provenance, of insufficient granularity, and incompatible with research protocols [35]

1.3 Semantic Interoperability

To fully realise the potential of EHR systems we need to ensure a timely and secure access to such systems to all those that are entitled to use them. Moreover, the information contained in EHRs should be up-to-date, accurate and, in its communication to another location, system or language it should be correctly understood. This is called interoperability and the most challenging part remains achieving semantic interoperability of EHR systems [92].

Semantic interoperability is the ability to communicate meaning, the ability for data shared by systems to be understood at the level of fully defined domain concepts [ISO/TS 18308:2002].

Stroetmann et al [92] distinguishes four levels of interoperability, two of them relating to semantic interoperability:

- Level 0: no interoperability at all.
- Level 1: technical and syntactical interoperability (no semantic interoperability). Systems in this level are capable to communicate data among them, but not its meaning. Complex data interfaces should be deployed to achieve data integration.
- Level 2: Partial semantic interoperability. Systems in this level are capable to understand the meaning of some of the communicated data. There are two orthogonal levels of partial semantic interoperability
 - Level 2a: unidirectional semantic interoperability
 - Level 2b: bidirectional semantic interoperability of meaningful fragments
- Level 3: full semantic interoperability, sharable context, seamless co-operability between systems.

The majority of current systems belong to level 1, using integration engines to communicate with external systems. Very partial semantic interoperability is achieved by using some standard coding systems, as ICD 9 CM [ICD 9 CM].

1.4 Ontologies

In a computational sense, an ontology is an explicit conceptualization of the entities of a domain [32]. An ontology is a special kind of information object or computational artefact which means to formally model the structure of a system [33]. It includes machine-interpretable definitions of basic concepts in the domain and relations among them [68]. It constitutes a way to model knowledge, so an ontology represents a knowledge model over a specific domain. As knowledge can refer to any context, ontologies can represent very different types of models at very different abstraction levels.

Some of the reasons to use ontologies are [25, 68, 97]:

- To explicitly represent domain knowledge and apply inference processes on it. This is a common use in the Artificial Intelligence field.
- To share common understanding of the structure of information among people or software agents.
- To enable reuse of domain knowledge, mainly about general theories.
- To make domain assumptions explicit which enable to change these assumptions easily if our knowledge about the domain changes. Hard-coding assumptions about the world in programming-language code makes these assumptions hard to find and understand but also hard to change.

- To separate domain knowledge from the operational knowledge. We can describe general tasks which can be implemented with different concrete elements, establish restrictions in the concepts used when collecting data, etc.
- To analyse domain knowledge. This is possible once a declarative specification of the terms is available.
- To enable data aggregation, improve search, and allow the detection of new associations among concepts.

There are an increasing number of available ontologies over the biomedical field, much of them freely available in Internet. Classic examples are the Foundational Model of Anatomy [FMA], the ontology edited by the National Cancer Institute [NCI], and the Gene Ontology [GO]. The National Center for Biomedical Ontology [66] maintains a repository of biomedical ontologies to enable their use, educate the scientific community about biomedical ontology and to collaborate with other groups [Bioportal].

Although is very common to find references to SNOMED CT [SCT] and UMLS [UMLS] as ontologies, is our opinion that cannot be considered as such.

1.5 Terminologies

The concept of terminology refers to the terms -words and multi-word expressions- that are used in a particular field. But under the concept of terminology we can find a great variety of linguistic resources, such as controlled vocabularies, encoding systems, lexicons, thesaurus etc. All these resources have in common to define necessary terms to express meaning. Some of these

resources can provide rules for expressing composition of terms, as well as other functionalities.

A terminology defines, by extension, a set of terms as well as many basic relationships between them. These relationships can be of very different type. Although the terms of a terminology refer to concepts, terminologies cannot be considered as true knowledge models.

Some intended purposes of terminologies are:

- To define the useful terms for referring to domain concepts.
- To establish lexical relationships between terms: synonymy, preferably, lexical variations, etc.
- To establish rules to compose valid expressions using available terms.
- To construct a vehicle to communicate meaning.

Some examples of terminologies are SNOMED CT [SCT], LOINC [LOINC] and UMLS [UMLS]. The different versions of the International Classification of Diseases [ICD] could be considered more as systems of codes than as terminologies.

1.5.1 SNOMED CT

SNOMED CT [SCT] (The Systematized Nomenclature of MEDicine Clinical Terms) is one of the most important examples of biomedical terminology. SNOMED CT is the most comprehensive, multilingual clinical healthcare terminology in the world. SNOMED CT contains over 400,000 concepts. A concept represents a unit of meaning and can have several associated descriptions, each one representing a human-readable term that describes the same meaning (i.e. a synonym). SNOMED CT relationships link concepts to other concepts whose meaning is related in some way. These relationships provide formal definitions and other properties of the concept [89]. SNOMED

CT concepts are organized in hierarchies. Within a hierarchy, concepts range from the more general to the more detailed. Related concepts in the hierarchy are linked using the *is_a* relationship.

Examples of some hierarchies include clinical finding, procedure, observable entity, body structure, organism, pharmaceutical/biologic product, substance and event.

1.6 Data Structures

A data structure is a way of organizing data [15]. A data structure is designed to organize data for a specific purpose. To achieve the communication of arbitrary health information between different EHR systems, a generic representation is required. This information architecture is called a reference model (RM).

A reference model is an information model that represents the global characteristics of the components of health records, how these components can be aggregated and the context information needed. A reference model defines both the structural elements and their relationship to communicate information. Some examples of reference models are CEN/ISO 13606 [EN 13606], openEHR [openEHR] and HL7 RIM [HL7 RIM].

Each one of the above standards defines its own artefacts for defining data structures types. CEN/ISO 13606 and openEHR call them archetypes and templates, and communicate clinical data as extracts. HL7 defines templates and communicate clinical data as CDA (Clinical Document Architecture) [CDA] documents.

Some uses of data structure types are:

- To establish models for collecting data.

- To establish restrictions in the data collection, according to knowledge models.
- To establish the useful relationships between instances of entities.
- To reflect the whole context where data are collected.

1.7 The CEN/ISO 13606 standard

The CEN/ISO 13606 standard [ISO 13606-1, ISO 13606-2, ISO 13606-3, EN 13606] is an standard composed by five parts which its overall goal is to define a rigorous and stable information architecture for communicating the EHR (all or part of it) of a single patient. The aim of the standard is to support the interoperability of systems that need to communicate EHR data while preserving the original clinical meaning intended by the author. Further aim is to reflect the confidentiality of that data as intended by the author and patient.

The approach adopted by this standard is named the ‘dual model approach’. It distinguishes a Reference Model (defined in Part 1), used to represent the global characteristics of health record components, and Archetypes (conforming to an Archetype Model, defined in Part 2), which are formal expressions representing a clinical recording scenario.

The Reference Model (13606 RM) comprises a small set of classes that define the generic building blocks (entities) to construct EHRs: folder, composition, section, entry, cluster and element.

The Archetype Model (13606 AM) comprises a set of classes to identify, define and describe archetypes, which are prescribed combinations of the building-block classes defined in the Reference Model.

An archetype is a structured and constrained combination of information entities which standardize information collected when describing instances of a given concept, such as blood pressure measurement.

Both the Reference Model and the Archetype Model (13606 AM) are defined in the standard as UML diagrams. The two models are completely independent; there are no common classes between them. Each kind of entity of the Reference Model used in an archetype is specified as the string-value of a property of the Archetype Model.

An archetype interchange format, called Archetype Definition Language (ADL), is proposed in Part 2 of the standard. Although ADL is optional in the standard, the Clinical Information Modeling Initiative (CIMI) [CIMI], an international collaboration dedicated to providing a common format for the representation of semantically interoperable health information, has chosen ADL as a formalism for representing clinical models.

1.8 The CEN/ISO 13940 standard

The purpose of CEN/ISO 13940 standard *Health Informatics - System of concepts to support continuity of care* [ISO 13940], is to define the generic concepts needed to achieve continuity of care, the concepts to represent both the content and context of the healthcare services.

The system of concepts defined in this standard is based upon the clinical perspective with the clinical process as focus, considering patient-centred continuity of care. This standard will establish a common conceptual framework across national, cultural and professional barriers.

This standard aims to be used whenever requirements for information in healthcare are specified, for example in the development of enterprise models, information systems and structured information for specified types of clinical processes.

Accordingly with the standard, to cover continuity of care, concepts are needed from all of these basic process aspects:

- Healthcare/clinical processes
- Management
- Support

The system of concepts defined by the CEN/ISO 13940 standard is based upon the healthcare/clinical processes. All other areas of work in healthcare both relate to and interact with the healthcare/clinical processes. As such, the management aspects and the resource support areas of healthcare are identified.

The part 1 of the standard divide the included concepts in the following areas:

- Concepts related to healthcare actors
- Concepts related to healthcare matters
- Concepts related to activities
- Concepts related to process
- Concepts related to healthcare planning
- Concepts related to time
- Concepts related to responsibilities
- Concepts related to information management

1.9 Structure of the Document

The rest of this document is organized as follows. Chapter 2 provides a state of the art in the areas related to this thesis. Chapter 3 defines the objectives, hypotheses, contributions, assumptions, and restrictions of this research. Chapter 4 presents an overview of the proposed architecture, OntoEHR, and the requirements the system is supposed to cover. The different components of the architecture are described in the following chapters, being the storage solution (OntoDDB) in Chapter 5, the framework for building applications (OntoCRF) in Chapter 6, the semantic interoperability approach (OntoCR) in Chapter 7 and the proposed EHR model in Chapter 8. The evaluation of the different components is presented in Chapter 9. Finally, the conclusions and future work are described in Chapter 10.

1.10 Derived Publications

During the development of this thesis the resulting achievements were published and presented to the scientific community after a peer review process. Such publications are gathered in this section.

Journal articles

- **Lozano-Rubí R**, Muñoz Carrero A, Serrano Balazote P, Pastor X.
OntoCR: A CEN/ISO-13606 clinical repository based on ontologies. J Biomed Inform 2016 Feb 18
doi: 10.1016/j.jbi.2016.02.007

- **Lozano-Rubí R**, Pastor X, Lozano E. OWLing Clinical Data Repositories with the Ontology Web Language. JMIR Med Inform 2014;2(2):e14
URL: <http://medinform.jmir.org/2014/2/e14/>
doi:10.2196/medinform.3023
- **Lozano Rubí, Raimundo**. Interoperabilidad semántica entre los sistemas de información sanitarios. Todo Hospital 2009; 258: pgs. 441-448
- **Lozano Rubí, Raimundo** ¿Cómo cuidamos la información en los hospitales? Todo Hospital 2006; 228: 384-390
- Pastor Duran, Xavier; **Lozano Rubí, Raimundo**. Sistemas de Información hospitalarios: barreras a superar. Todo Hospital 2005; 218: 385-393
- Geva Urbano, Felipe; Saiz Codina, Silvia; Pastor Durán, Xavier; **Lozano Rubí, Raimundo**. Representación del conocimiento de INBIOMED. Sociedad Española de Informática y Salud 2004, 46:18-21

Conference papers

- **Lozano-Rubí R**. Revisión de la norma ISO 13940. IV Reunión del Foro de Interoperabilidad en Salud. Valencia. 2014
- **Lozano-Rubí R** . Aplicación de la terminología SNOMED CT en el registro de cáncer de mama y la codificación de medicamentos. X Jornada del Fórum Catalá d'Informació i Salut. Barcelona. 2012

- **R. Lozano-Rubi**, X. Pastor. Building a Clinical Repository based on ISO-13606. Proc Third Symp Healthcare Systems Interop, 2011,ISSN 2174-7415
 - **R. Lozano-Rubi**, X. Pastor. The Puzzle of Semantic Interoperability. With regard to OntoCRF. Proc Second Symp Healthcare Systems Interop, 2010, ISSN 2174-7415, pp 57-66
 - **R. Lozano**, X. Pastor, and E. Lozano. OntoDDB - ONTOLOGY DRIVEN DATA BASE. Proc First Symp Healthcare Systems Interop, 2009, ISSN 2174-7415, pp 31-38
- Lozano-Rubi R**, Pastor X, Lozano E. OntoDDB - Ontology Driven Database. Proc First Symp Healthcare Systems Interop; 2009: 31-38, ISSN 2174-7415
- **Lozano-Rubi R**. La interoperabilidad semántica. VII Jornada del Fórum Catalá d'Informació i Salut. Barcelona. 2009
 - **Lozano Rubí, Raimundo**; Pastor Duran, Xavier; Urbón-Bayés, Pere; Lozano Hontecillas, Esther. OntoDDB – Ontology Driven Data Base. Libro de comunicaciones y pósters de Infors@lud 2008. XI Congreso Nacional de Informática de la Salud. Abril 2008. ISBN: 978-84-691-2468-0. pgs. 273-279
DOI: 10.13140/2.1.3749.3768
 - **Lozano, R.**; Pastor, X.; Urbón,P.; Lozano, E. OntoDDB – Ontology Driven Data Base. VI Jornada del Fórum Catalá d'Informació i Salut. Barcelona. 2008

- **Lozano, R.** De los datos al conocimiento. INFORMED 2004 - X Congreso Nacional de Informática Médica. Barcelona. 2004
- **Lozano, R.** Compartir Coneixement mitjançant ontologies. II Jornada del Fórum Català d'Informació i Salut. Barcelona. 2004
- **Lozano-Rubi R,** Geva F, Saiz S, Pastor X: Relational Support for Protégé. Sixth International Protégé Workshop; 2003. Manchester.

Conference posters

- **Lozano-Rubí R;** Pastor-Duran X; Muñoz A; Serrano P; Conesa A. Standard Interchange of Clinical Data Between Health Organizations. WHO-FIC Network Annual Meeting 2015. Manchester
- **Lozano-Rubí R;** Pastor-Duran X; Conesa A. Linking clinical data with multiple terminologies. WHO-FIC Network Annual Meeting 2014. Barcelona
- **Lozano-Rubí R,** Pastor-Duran X, Canela-Soler J. Development of a native 13606 clinical repository using a conceptual approach. WHO-FIC Network Annual Meeting 2013. Beijing, People's Republic of China.

- **Lozano R**, Geva F, Saiz S, Pastor X. Structuring Medical Knowledge. MIE2003 – XVIIIth International Congress of the European Federation for Medical Informatics. Saint-Malo (Francia)

Book chapters

- Pastor Durán, Xavier; **Lozano Rubí, Raimundo**. Representación de la información clínica e investigación sobre salud. Informática Biomédica 2004; INBIOMED; ISBN: 84-609-1770-3; pg 115 y ss

1.11 Patents

It is important to mention that the intellectual property of the software products obtained during this thesis have been protected, Spanish laws not allowing patents over software.

- Inventors: **Lozano Rubí, Raimundo**; Pastor Durán, Xavier
 Title: OntoDDB: Sistema de recogida y almacenamiento de datos dirigidos por ontologías
 Entry number recorded: 02/2009/6941
 Application number: B-3774-09
 Country: España
 Date of presentation and effect: 13/07/2009
 Owner entity: Hospital Clínic de Barcelona y Universidad de Barcelona

- Inventors: **Lozano Rubí, Raimundo**; Pastor Durán, Xavier
Title: OntoCRF: Onto Clinical Research Forms
Entry number recorded: BC 7847671
Application number: ADP1638
Country: España
Date of presentation and effect: 29/12/2012
Owner entity: Hospital Clínic de Barcelona y Universidad de Barcelona
Empresa/s que la están explotando: COSTAISA

- Inventors: **Lozano Rubí, Raimundo**; Pastor Durán, Xavier
Title: OntoCR: Onto Clínic Repository
Entry number recorded: BD 1126007
Application number: ADP89
Country: España
Date of presentation and effect: 22/01/2013
Owner entity: Hospital Clínic de Barcelona y Universidad de Barcelona

Chapter 2

State of the Art

Healthcare is provided by a network of healthcare facilities belonging to different organizations distributed over the territory, characterized by a high degree of heterogeneity. A large number of software applications for health care, mutually isolated and incompatible are already available on the market. Even within the same centre, healthcare information systems are frequently fragmented across a number of applications isolated and scarcely consistent with each other. Whatever healthcare organization can be described, according to a federated model, as a set of organizational units which mutually interact to provide services. Each organizational has an assigned mission and responsibility, which result in its own business rules, implemented in specific software systems. Hence, each organizational unit has some autonomy and independence, in terms of information managed and activities performed.

On the other hand, increasing longevity and complex treatments mainly related with chronic conditions, led to an increasing tendency to share the patient care among the different healthcare levels. This increases the degree of heterogeneity of information systems used to provide a view of the patients' health state; a view which should be unique and coherent.

2.1 EHR systems

EHR systems are used in very different ways [31] and most healthcare institutions have some form of EHR in place, following different approaches in their development of an EHR.

As a consequence of the variety of organizations providing healthcare and the heterogeneity of information systems used, current EHR systems are not capable to show to healthcare professionals a conceptually consolidated view of the patients' health state.

Building systems capable to work in a distributed environment is not an easy task. It does not seem feasible to establish a single technical solution that meets all the requirements, so it is necessary to establish an open and flexible architecture that ensures functional and semantic interoperability. This objective can be achieved through the use of standards, so different providers could collaborate in deploying distributed systems. In recent years, international bodies of standardization have developed several standards in this direction.

The standard CEN/ISO 12967 [ISO 12967] proposes an open architecture based on a middleware layer, which defines a set of workflows, information and services common to all healthcare information systems, with the aim to allow technical and functional interoperability between the different systems, modules and components.

The use of standard CEN/ISO 13606 [ISO 13606], together with the use of classification systems and standard terminologies, provide the mechanism to assure semantic interoperability when communicating clinical information among information systems.

Finally, the standard CEN/ISO 13940 [ISO 13940] defines the generic concepts needed to achieve continuity of care, with the focus on the clinical process, and with a vision centered on the patient.

Each of the three standards mentioned above covers different aspects. The concurrent use of the three of them could facilitate overcoming current barriers.

In recent years, governments of industrialized countries have promoted the development of national programs for healthcare information technology. The approach followed and the results obtained have been different [28]:

- The top-down approach: development of nation-scale health information systems to create a single record. It is the case of Connecting for Health [NHS CFH] in United Kingdom. The results obtained in this project

were far from expected and the project was shut down on 31st March 2013.

- The bottom-up approach: to interconnect existing systems into health information exchanges (HIE). It is the case of the United States. The HIE approach does not create a single record, but intends to allow virtual views of records, as abstracted or aggregated from regional systems. The expectation is that Regional HIEs will eventually aggregate into a nation-scale system. Despite the EHR adoption has grown remarkably over the last years, and the evidence of benefits in safety and quality, many clinicians claim that EHR use has had unintended clinical consequences [74].
- The middle-out approach: creating a common set of technical goals, standards development, and support for standards implementation. The example is the Australia's project and the Australia's national E-health transition authority (NEHTA) [NEHTA], whose purpose is to define the interoperability standards that will be used to specify any future national health information system.

When analyzing the various dimensions that current and past approaches to building electronic health record systems have addressed, semantic interoperability is the most significant challenge still to be solved [13].

2.2 Semantic Interoperability

Semantic interoperability is an essential factor in achieving benefits from EHR systems to improve the quality and safety of patient care, public health, clinical research, and health service management [92].

Basically, semantic interoperability between two agents is based on the following premises:

1. The agents are able to communicate each other. Sent messages arrive efficiently to each other and the receiver can syntactically process the message, being able to decode it. This only would assure technical interoperability.
2. The agents use a common language; share a vocabulary to refer to the concepts of the domain.
3. The agents talk about the same things; share a common domain of concepts.

So far, only an efficient implementation of the first premise has been achieved. As commented in section 1.3, current systems in use have only capabilities for technical interoperability. These capabilities rest on the use of a message system and an integration engine which code and decode the messages.

The Open Systems Interconnection model (OSI model) [ISO-OSI] is a conceptual model which establishes seven different levels of communication, the functionality of each level and the interfaces between them. A set of standards (IP, IPX, TCP, UDP, SPX, NetBEUI, NFS, SQL, RPC, JPEG, MIDI, MPEG, MP3, SNMP, FTP, TELNET, HTTP, HL7) allow the implementation of the required functionality at each level.

The division into levels and the mentioned used standards have allowed the emission of a message by an application and the correct reception by another one, which is able to access the content of the message. But in any case the receiver application can understand the whole content of the message. For doing so, it is necessary a common set of concepts and a common vocabulary for referring them.

In the last years, there has been a great development of standards and norms to guarantee semantic interoperability between eHealth systems, being the most

important: HL7 V3 RIM and CDA, CEN/ISO 13606, openEHR, CEN/ISO 13940, and SNOMED CT, some of them described in chapter 1. Moreover, a considerable amount of effort regarding international harmonization is underway [6]. Countries known to have achieved high levels of EHR use, as Denmark, New Zealand, and Sweden, have national-level policies about interoperability standards [30].

Generic information models, clinical models, ontologies and terminologies have been identified as required artifacts to achieve semantic interoperability [57], but closer integration between these elements is needed [36, 92, SemanticHealthNet]. It is not well understood what aspects can solve each standard. Because of it, there is not a proposed architecture for defining different levels, functionalities and responsibilities, needful for a perfect definition of each standard or protocol, and their integration with the rest. The consequence can be a confuse application of the different standards, which can overlap among them in some areas.

In recent years, research work has been done on both technical transformation between different approaches proposed [44, 58] and the construction of standardized repositories for secondary use of clinical data [27, 54, 73, 79]. However, there are very few proposals for real implementations of interoperable EHR systems for primary use.

2.3 Research support

The Hospital Clínic of Barcelona has an EPR system since 1995. Three different commercial systems have been used during this time, the last one including a data warehouse, but they were mainly focused on economic and administrative

processes. Although these systems allowed gathering some limited clinical data, any of them were intended to register additional data.

Financial limitations results in a widespread use by researchers of basic and general-purpose applications, such as spreadsheets. The same situation is reported by other authors [3, 37]. The use of general-purpose applications have serious drawbacks: not friendly user interface, few guarantees for maintaining the consistency of data, difficulties for sharing and consolidation of data, and limited ability to exploit data. Desktop applications are definitively not designed to meet these criteria.

When an adequate budget is available, it is possible to build a more sophisticated system. Usually, these systems are built using a multi-tier architecture composed by a centralized database, an application server, and a web server providing the user interface. However, this architecture presents some disadvantages. First of all, the development of such applications is a laborious task, as so is their extension to accommodate changes. Consequently, this approach is not suitable for domains where data and model evolution is the norm [47]. Secondly, this classical approach requires a very specialized panel of computer technicians which often leads to communication problems between the biomedical researchers and the development team. Thirdly, the return of investment is very low. The development cost and the cost of IT personnel, suppose a high investment [3], sometimes for a very short period of time; research projects have a typical extent of 2-3 years. And last but not least, this kind of approach in the bosom of an organization produces a big heterogeneity among the different applications devoted to research projects, and a distribution of data across multiple sources, which complicates the ability of researchers to use the data for answering their research questions [27].

2.4 Use of ontologies in EHR systems

OntoEHR proposes the use of ontologies to declaratively define the metamodel which drives the system. The ontology-driven development of complex and intelligent systems has been largely applied in the past, especially when the ontologies or the methods are likely to be reused for new or derivative applications [40, 63, 65]. In general, the goal is to transform the system development cycle, so instead of programming each new application from scratch, we can select, modify, and assemble existing components [64]. Ontologies are used to build knowledge bases containing detailed descriptions of particular application areas. The proposed model goes a step further, there is no need of any kind of programming, just the design of the application ontology. In OntoEHR ontologies contain not only knowledge about the domain, but also the detailed description of the application.

The discussed approach is also related to the *Model-Driven Architecture* (MDA), launched by the Object Management Group (OMG) [17]. According to their manifesto, MDA is a style of enterprise application development and integration, based on using automated tools to build system-independent models and transform them into efficient implementations. As with ontology-oriented approaches, software evolution is handled simply by editing the underlying model. OMG is guided to object oriented applications, particularly to distributed ones. It represents a more technical approach, centred on the platform independence, whereas the model we propose pursues the conceptual independence. In our case, the database never changes; neither does the implementation of the application. On one hand, our work represents an advance since everything is defined explicitly. On the other hand, the use is much more restricted.

In regard to research in data repositories, there exist multiple ontology-driven solutions for discovering and searching existing resources [Datacite, eagle-i], or to consolidate clinical research data from disparate databases [27], but not so much for automatically building new ones.

Compared with Protégé, WebProtégé [96] adds collaboration support and improves knowledge acquisition, but remains mainly an ontology editor. The work of Li et al [47] is close to our work in considering ontologies as the centre of the architecture. The proposed system is focused to model a domain and to support data and model changes, through versioning and dynamic composition, while using a simple interface with few options. On the other hand, Butt et al [22] propose the automatic generation of web forms from ontologies with the objective of facilitate the creation of RDF data. While the system produces forms easy to use, the capabilities of structuring the information are very limited.

At the moment of writing and at the best of our knowledge, there are no frameworks allowing the creation of data repositories, with the interface functionalities of traditional systems, in such a dynamic way like OntoEHR, where even the user interface is built through the edition of ontologies.

In the particular field of storage of RDF graphs and ontologies, there exist in the bibliography several works like triple stores and relational databases with RDF-based access. However, none of these fulfil the needs of our system, since they do not take into account the semantics of ontologies. In the case of RDF-based access to relational databases, like the platform D2RQ [D2RQ], the system is read-only and just provides a "RDF view" of the content, but it does not provide any solution for storing the content, relying on an existing database created by the user. Also, the user has to generate the mappings between the platform and the database, specific for each use case. In the case of triple stores, they offer a way to store and retrieve triples, leaving for an API or for a query engine the

logic necessary for interpreting the triples and retrieving the right ones. This requires to analyse the whole set of triples of a specific graph, which has a high cost and prevents the system to scale.

Nevertheless, our repository is not the only one with these characteristics, although it was at the time of searching for solutions. Systems like OWLIM [14] and DLDB2 [71] combine database management systems with additional capabilities for partial OWL reasoning. Furthermore, there are repositories with similar architectures to ours, like Minerva [103] repository within the Integrated Ontology Development Toolkit by IBM. Since OntoEHR has not any SPARQL capability yet, we could not perform any reliable comparison under equivalent conditions to these similar repositories.

2.5 Summary

In this chapter we have established the context of this thesis regarding existing works in the main areas of our research. Firstly, we have reviewed current approaches to build semantically interoperable EHR systems. Although the more successful systems are those based on the use of international standards, the degree of semantic interoperability achieved in real systems is very limited. Secondly, regarding research support the use of general-purpose applications, such as spreadsheets, is common. The alternative is to build costly applications that cannot accommodate changes easily.

Finally, we have analysed some approaches for the development of information systems enabling more independence and reuse of different components. Whilst the MDA initiative is centred on the platform independence and existing ontology-driven proposals are centred on reusing conceptual components, our

approach moves all system specification to the conceptual level, achieving both objectives.

Chapter 3

Goals and Contributions

This chapter describes the goals of this thesis and the resulting contributions, along with the assumptions, hypotheses and restrictions that made such contributions possible.

Our vision is to provide true support to healthcare professionals in their daily practice, for both primary and secondary use of clinical data.

3.1 Objectives

The main goal of this thesis is to put the basis for a new EHR model driven by ontologies, through the development of a metamodel to integrate clinical data coming from heterogeneous sources. In order to achieve this general objective, it is necessary to establish a list of more specific goals to be accomplished.

Specific objectives:

- O1.** To assist healthcare professionals gathering and recording structured clinical information and its reuse, as electronic medical record, epidemiological registries and clinical research.
- O2.** To facilitate the continuity of care between health facilities.
- O3.** To design a system that is at the same time flexible, solid, and efficient, capable to deal with the complexity and change of clinical domain.

O4. To seamlessly incorporate clinical knowledge into the clinical information systems.

3.2 Contributions

To achieve the aforementioned goals, in this thesis we generated the following main contributions:

- C1.** OntoEHR, a conceptual architecture for a new semantically interoperable Electronic Health Record system driven by ontologies.
- C2.** OntoDDB, a framework for the definition and storage of ontologies, which can be used to build both knowledge servers and data repositories.
- C3.** OntoCRF, a framework for the definition, modeling, and instantiation of data repositories, including the storage, the specification of the application to manage the data and the GUI.
- C4.** OntoCR, a semantically interoperable clinical repository, based on ontologies, and conforming to CEN/ISO 13606 standard.
- C5.** A Problem Oriented Medical Record model, focused on the representation of the clinical process.
- C6.** A tool that implements the above models, thus allowing the recording of clinical data in a new EHR system.

3.3 Hypothesis

The contributions of this research rely on several initial hypotheses. We enumerated and classified them as general or specific ones, as described in the following subsections.

3.3.1 General hypothesis

The general hypotheses of this thesis are related with the main objectives to be achieved.

- Representing a better model of the patient. Representing the clinical process in healthcare information systems will provide better support to healthcare professionals. Using existing international standards to do it, add semantic interoperability.
- Ontologies can be successfully applied to explicitly represent the conceptual layer of EHR systems, and drive the whole behavior of the system.

3.3.2 Specific hypothesis

We divided the general hypotheses into the following specific hypotheses, which could be feasibly evaluated:

H1. A relational database designed following the OWL model is suitable for ontology storage and edition.

H2. Ontologies can be used to semantically integrate specific clinical data.

H3. Using an ontology-based approach it is possible to build applications addressed to health professionals.

H4. Standard archetypes can be used to build clinical applications.

H5. Modeling clinical information using ontologies, archetypes and controlled vocabularies is a suitable method to communicate clinical information between healthcare settings maintaining the semantic of the information.

3.4 Assumptions

The following assumptions were considered in this thesis:

A1. We assume the definition of clinical process established by CEN/ISO 13940:2007 standard as the core of the system.

A2. An important characteristic of the system proposed in this thesis is to integrate concrete clinical data with ontologies. We assume biomedical ontologies already exist and are maintained by third parts.

A3. This thesis relies on existing CEN/ISO 13606 archetypes defined by an authority organization responsible for their governance.

A4. Management and resource areas are not considered in this work (see **R3**). Nevertheless, they constitute an essential part of healthcare systems that we assumed is already implemented in some ERP system. Therefore connections between them must be considered.

3.5 Restrictions

In this thesis, there are some restrictions that define the limits of our contribution and may establish future research objectives. These restrictions are the following:

- R1.** All ontologies in this thesis are represented in OWL 1 language. The tools implemented in this work cannot be used with ontologies in OWL 2 format.
- R2.** It is out of the scope of the system to incorporate or to generate data using other EHR models of reference than CEN/ISO 13606.
- R3.** In this thesis healthcare services are seen from a medical/clinical perspective. The system proposed is focused on the clinical process as is described by the standard CEN/ISO 13940:2015. Management and resource support areas are not objectives of this thesis.

3.6 Research Methodology

In this research we followed a requirement-driven approach with empirical validation. From the beginning, the main goal was to design a standardized model of shared electronic health record (EHR) to be implemented in a real scenario. With this final picture in mind, we advanced in a step-wise approach, by successive phases, using real use cases to define requirements and validate the solutions. Our work in the Hospital Clinic of Barcelona allowed us to define the goals, requirements and hypotheses to validate. The functional requirements an EHR system should cover and the POMR proposed, are mainly fruit of this experience and the interaction with healthcare professionals. In addition, we reviewed the state of the art of related approaches as presented in Chapter 2. On the other hand, the majority of hypotheses were validated building real applications. We have applied a progressive and iterative research methodology, adding tools and functionalities incrementally:

- First focus was the design and implementation of a relational repository of ontologies, with the objective of separating the physical storage of data

from the conceptual structure of these data, to gain flexibility in front of new requirements and modifications.

- Second step was to extend the functionality of Protégé to connect it with the ontology repository and to start using the system to both to design the data model and to store specific data. In this way, we used Protégé as a kind of database design tool. Data were stored in a relational database and accessed from the ontology editor (Protégé).
- Third goal was to prove the capability of such a system to integrate heterogeneous data under a conceptual framework. So, the meaning of codes do not remain implicit in the database but was declaratively stated in the ontology.
- Follow on, the next step was to extend the system to develop a framework for building clinical repositories and applications for research: OntoCRF. With OntoCRF, is possible not only to build data repositories but entire applications. The GUI of the application and the main behaviour is declared in the ontology. This is achieved adding a new layer on top of the system, a metamodel represented also by an ontology.
- The metamodel of OntoCRF was then extended to incorporate the structure of CEN/ISO 13606 standard, ISO 21090 standard and SNOMED CT. This new way of structuring the data is achieved maintaining the original structure, as adding a new way of viewing the same data. The result is OntoCR, a standardized clinical repository with an invariant relational storage and a variant metamodel which drive all the system and can cope with change and evolution.
- The last step was to design a new EHR system incorporating a POMR model.

3.7 Summary

This chapter describes the objectives, contributions, hypotheses, restrictions and assumptions taken into account in this thesis. Table 3.1 provides a summarized view, making explicit the relations among them. Finally, in this chapter we describe the research methodology followed in this thesis.

Objective	Contributions, hypotheses, assumptions, and restrictions
O1. To assist healthcare professionals gathering and recording structured clinical information and its reuse, as electronic medical record, epidemiological registries and clinical research	Contributions C1, C3, C5, C6 Hypotheses H2, H3 Assumptions A1, A3, A4 Restrictions R2, R3
O2. To facilitate the continuity of care between health facilities	Contributions C1, C4, C5 Hypotheses H2, H4, H5 Assumptions A1, A3 Restrictions R2, R3
O3. To design a system that is at the same time flexible, solid, and efficient, capable to deal with the complexity and change of clinical domain	Contributions C2, C3 Hypotheses H1, H2, H3 Assumptions Restrictions R1
O4. To seamless incorporate clinical knowledge into the clinical information systems	Contributions C1, C2, C3 Hypotheses H1, H3 Assumptions A2 Restrictions R1

Table 3.1 – Relation between the objectives and their corresponding contributions, hypotheses, assumptions, and restrictions.

Chapter 4

OntoEHR - The General Model

Traditional EHR systems currently in use have been based on financial and administrative activities. Administrative processes and B2B workflows are generally well represented but, as previously mentioned, current systems fail to provide true support to healthcare professionals. Although clinical data has been translated from paper records to databases, the result obtained, in a majority of cases, is a kind of electronic paper, with clinical processes not represented at all. For example, clinical relationships between data are not present in an explicit way. This approach, OntoEHR, aims to define a conceptual architecture for a new semantically interoperable Electronic Health Record (EHR) on the basis of representing the clinical process.

To support the clinical process a system must be able of properly representing health states of patients, and their links with medical knowledge. For this reason, the core of the system we propose in this thesis is a standardized clinical repository integrated with knowledge services.

OntoEHR does not pretend to substitute systems currently in use but to empower them with true clinical value. Clinical processes are dependent upon management and resource. These aspects of healthcare are already well supported by current ERP systems and are not objectives for OntoEHR.

Integration will be needed between OntoEHR and an ERP system to seamlessly relate these areas between them.

4.1 Functional requirements to cover

The Hospital Clínic of Barcelona is a centennial institution with a wide implantation of different computerized systems for the collection of clinical data during last decades. Planning activities, registering activity, generating clinical reports, service request, drugs prescription or diagnose coding are supported by information systems resulting in an extensive and intensive use, with high implication of health personnel.. On the other hand, the HCB has a long tradition in biomedical research, exploiting recorded data, as well as conducting prospective studies. . The unit of Medical Informatics at Hospital Clínic of Barcelona has a long experience designing and helping in the implementation of these systems, both for supporting daily clinical practice and research [48, 49, 72].

In order to establish a set of requirements for OntoEHR we distinguish among three groups. Firstly, the general requirements about clinical data management reported in the literature [93]. Secondly, innovative capabilities of the system, fruit of the experience at the HCB, that we consider essential.. A third set of capabilities is related with research needs. Although some of these research capabilities are implicit in the other two sets, we have preferred to maintain this little group aside because its specificity.

4.1.1 Essential capabilities

We consider as essential capabilities of the system those which constitute a differential characteristic of the proposed model regarding traditional models in use.

These capabilities are as follow:

- **Patient centered.** System centered on the patients and their health problems. This implies a global view of the patients' health data that can be extended beyond the limits of a unique organization. The proposed system does not impose organizational limits, so can be used by different organizations, provided some requirements (mainly the use of some standards) are accomplished. Information related patients' health problems and medication plan are clear examples of data whose management are carried out by professionals spread among several organizations, and must be globally viewed, as a consolidated picture composed by the operation of different information systems.
- **Patients' health problems as the core of clinical management.** It is the real business of the healthcare professionals. A proper management until their resolution is what the patient (customer) expects from physician decisions. The simple record of patients' health problems is not enough. The system should provide support for their entire management.
- **Health professional oriented.** The system should be oriented towards the healthcare professional supporting all the actions that eases the decision support, with an active participation of the patient and his environment. The system aims to provide support to the clinical process, so the interaction of the different participating agents with the system is crucial, mainly healthcare professionals and patients. This tool should be reliable, useful, and provide true added value to the healthcare professional and the patient.
- **Knowledge representation.** Integration among data and knowledge coming from different sources. Knowledge of varied nature is a main resource of the clinical process, and should be integrated with available data. Thanks to TIC evolution, nowadays there is knowledge available in different formalized formats, a rising tendency.

- **Knowledge driven.** To access to available knowledge in query mode is not enough in the medical domain. Existing knowledge models should be used to really drive the system. Integrating knowledge models could provide taking automatic decisions in an autonomous mode, assisting healthcare professionals suggesting what to do, and providing available evidence.
-
- **Semantic interoperability.** Full semantic interoperability with other systems. The clinical repository should be able to communicate clinical information to other systems using CEN/ISO 13606 extracts to ensure the extension of the integrated view over the patient, with the contribution of other systems.

4.1.2 General capabilities

Following the US IOM report, Key Capabilities of an Electronic Health Record System [93], whatever clinical information system implemented on the proposed architecture should satisfy some general requirements. These capabilities should be considered in order to guarantee that the model proposed in this thesis does not represent any obstacle to achieve them.

- Availability of stored clinical data for querying and updating in real time, such as patients' diagnoses, allergies, lab test results, and medications. This availability should be permanent and ubiquitous, when and where is needed.
- Integration with order management. The existence of a health problem constitutes the trigger of the process, and all subsequent healthcare activities should be related with it. The clinical repository should contain the necessary links with management elements.
- Electronic communication and connectivity. Efficient, secure, and readily accessible communication among health care team members and patients. The system should provide on-line communication; expanding the

typology of encounters and activities among healthcare professionals and patients.

- Electronic data storage should employ uniform coding standards for clinical, organizational and management purposes.
- Patient support. Tools that give patients access to their health records, provide interactive patient education, and help them carry out home-monitoring and self-testing.
- Capability to define roles to limit the data accessibility, according to well defined criteria.
- Security and traceability of data access. The management of clinical data should be conform to specific dispositions in terms of security, validation and traceability.
- Data exploitation. Data stored in the clinical repository should respond to different reporting requirements, including those that support patient care, disease surveillance, clinical research, and management.

4.1.3 Research requirements

In the context of clinical research, the main general requirements about data management reported in the literature [37, 54, 47] are as follow:

- The ability to efficiently acquire, store and manage large volumes of structured data, preferably in a centralized repository.
- To provide a web interface for researchers to allow them to have a distributed access to the data, in order to introduce new data or to retrieve existing data. Data are usually gathered by various researchers, often in different locations.
- Data security, including access control, to assure the persistence of the data.
- To facilitate the access to the data, including researcher ‘self-service’ access.

- To be able to easily accommodate changes in the structure of the data, minimizing service disruption when such a model change occurs.

4.2 Workflow

The standard CEN/ISO 12967 [ISO 12967] identifies three fundamental workflows in the users' activities:

- Subject of Care workflow (patient-centric)
- Activity management workflow (carer-centric)
- Clinical information workflow (information-centric)

This proposal refers to the clinical information workflow. This workflow relates to users' activities related to the management of the clinical data. Management of classifications, coding criteria and dictionaries adopted in the different sectors to classify the managed information, is also considered in this proposal.

4.3 OntoEHR Overview

As stated previously, OntoEHR represents the clinical process, as the set of interrelated healthcare activities which improve or maintain the initial health state of a patient [ISO 13940].

In this thesis, we do not propose a concrete EHR system, but the foundations to build EHR systems focused on the clinical process. To achieve this goal we propose, basically, a data storage where work data are structured and persisted according a problem oriented model of the clinical process. All specifications of the system are declaratively defined in a metamodel by means of OWL ontologies. This approach guarantees technical independence and facilitates the evolution of the system.

The first component is the clinical repository, a database where concrete data about patients will be recorded. The second component is a framework to build the applications that will access and manage the recorded data. The third component is the deployment of semantic interoperability is a recognized need for this kind of systems, as stated in chapter 2. The clinical repository of OntoEHR is able to communicate clinical information with other systems in a standardized way. Biomedical knowledge (representation and operation) is the fourth constituent. The activities performed by healthcare professionals are driven by medical knowledge, an essential element that should be integrated into the system too.

These four elements: the clinical repository, the framework to build applications, the semantic interoperability capabilities and the knowledge server, constitutes the technical layer of OntoEHR.

A Problem Oriented Medical Record (POMR) model is proposed to structure the clinical data that will be managed in OntoEHR. We use the CEN/ISO 13940 [ISO 13940] standard as a general conceptual framework that describes the clinical process and the organizational concepts we need. Finally, the system architecture is inspired by the ISO 12967 standard [ISO 12967]. According to this one, the access to the different components should be provided by an independent middleware layer through the required services.

Chapter 1 has introduced the concepts of ontology, terminology, information model, and data structure. Although several examples have been shown pertaining to each category, frequently the different resources may cover, with different degree, more than one, causing overlaps and confusions.

Table 4.1 represents the class each of the most common resources belong to and shows, in our opinion, to what extent the phenomenon of overlapping occurs:

	Knowledge model	Terminology	Information model	Data structure
Ontologies	XXX	XX	X	
SNOMED CT	X	XXX	X	
UMLS		XXX		
RIM HL7	X		XXX	
CEN 13606			XXX	
OpenEHR			XXX	
CDA			XX	XXX
Archetypes	XX		XX	XXX
CEN 13940	XX	X	XXX	

Table 4.1 – Different kinds of resources and its intended use

Some of these elements are present in more than one column. Reference models and data structures are themselves a kind of information model. Whilst this situation does not cause any problem, the consideration of some resources as knowledge models, terminologies, and information models at the same time is confusing.

Therefore, it is very important to become aware of these overlapping and to separate assigned functionalities to the different available resources [83].

The first consideration to do is about the role of ontologies. Ontologies are the intended artefacts to represent knowledge models. So, in an integrated system, all the knowledge should be modelled explicitly using ontologies. This is the role assigned in OntoEHR to ontologies.

The second consideration is about terminologies. Vocabularies and terminologies, SNOMED CT in particular, should be used only for the purpose of identifying communicated data. These vocabularies and terminologies should be properly linked with ontologies. In OntoEHR all concepts should be identified by at least a terminology, SNOMED CT when possible.

Finally, information models and its derived structures do not constitute knowledge models or terminologies. They define the structural level but not the conceptual one. They should be used to implement at the syntactic level the restrictions defined at the conceptual level, plus the elements guiding the technical implementations. In OntoEHR, CEN/ISO 13606 extracts and archetypes are used to build forms to record data and documents to communicate them.

4.4 System Architecture

This thesis proposes a conceptual architecture for a new kind of EHR. The architecture of the OntoEHR system is shown in Figure 4.1 together with the rest of elements which are usually found around an EHR system.

The components of the system are divided in four layers. At the bottom, the persistence layer deals with the storage of the different kind of data: financial and administrative data to be used in ERPs for patient management; and clinical

data, research data and knowledge data to be used in OntoEHR. The application layer includes all the necessary logic of the different applications. At the top of the figure, a web portal constitutes the graphical user interface (GUI) both for professionals and patients. Finally, a service layer between the applications and the web portal allows the communication between the different components.

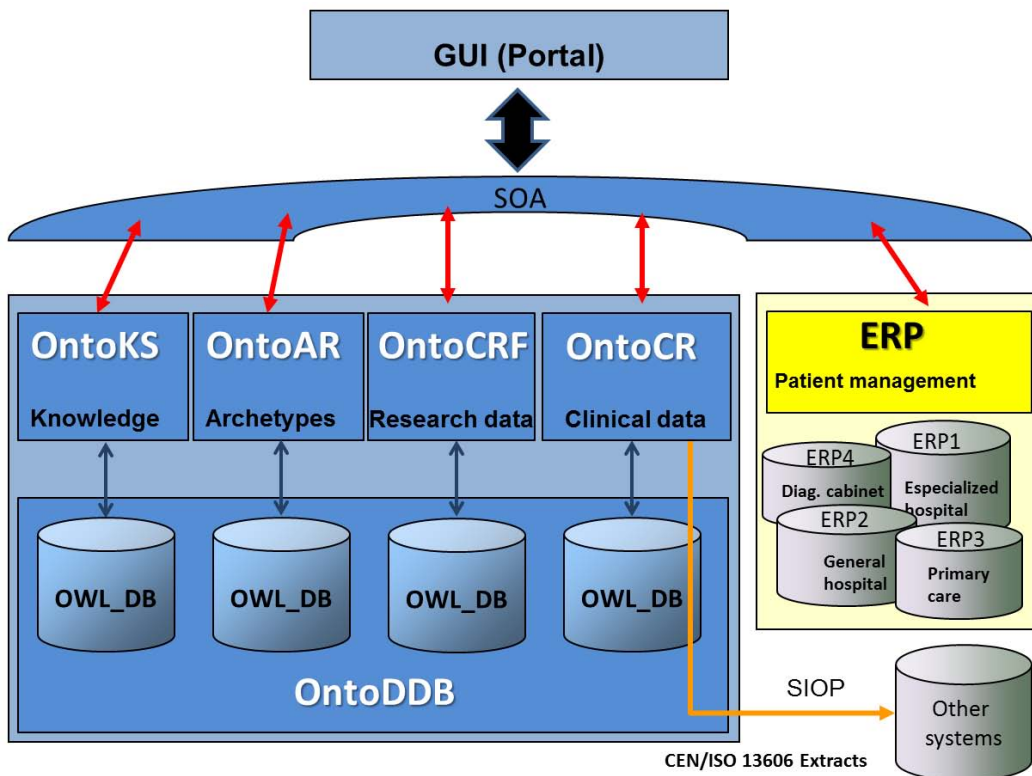


Figure 4.1 – General schema of OntoEHR

Notice that in Figure 4.1 the components of OntoEHR are represented in blue. The box in yellow represent different ERP systems which are not part of OntoEHR. Semantic interoperability with other systems is achieved by communicating CEN/ISO 13606 extracts.

The purpose of each component in the architecture is described as follows:

- OWL_DB (OWL Database) is a relational database used for storing ontologies and instantiated data. OWL-DB is a component of OntoDDB and constitutes the persistence layer of OntoEHR. It is not relevant if it is used a unique database or several databases dedicated to specific purposes. OWL_DB is explained in detail in Chapter 5.
- OntoDDB (Ontology Driven Data Base) is a framework for the definition and storage of ontologies. Using an external editor, like Protégé, allows the storage and retrieval of ontologies in a relational support. OntoDDB constitutes the editing tool of OntoEHR. OntoDDB is described in Chapter 5.
- OntoCRF (Onto Clinical Research Forms) is a framework for the definition, modeling, and instantiation of data repositories, covering the storage, the application to manage the data and the GUI. All the required information to define a new project is explicitly stated in ontologies. Moreover, the user interface is built automatically on the fly as web pages, whereas data are stored in OWL_DB. OntoCRF is described in Chapter 6.
- OntoCR (Onto Clinical Repository) is a semantically interoperable clinical repository, based on ontologies, conforming to CEN/ISO 13606 standard. OntoCR is an evolution of OntoCRF, a kind of standardized OntoCRF. Data managed in OntoCR are stored in OWL_DB. OntoCR is described in Chapter 7.
- OntoKS (Onto Knowledge Server) is the knowledge server for all the system. OntoKS uses OWL_DB as storage system and is intended to manage all needed ontologies, including, but not limited to, ontologies defining clinical knowledge, information about drugs, protocols,

guidelines, etc. OntoKS has capabilities to upload and download ontologies in OWL format.

- OntoAS (Onto Archetype Server). As archetypes are defined in OntoEHR in OWL format, archetypes will be managed by OntoAS, a kind of dedicated OntoKS.
- The GUI is defined in each subsystem (OntoCR, OntoKS, OntoAS) and deployed in Liferay [Liferay].
- Following the recommendations of ISO 12967 standard [ISO 12967], the implementation of HISA services can be described as a Service Oriented Architecture (SOA), e.g. in the form of web services. This thesis does not define the required services, which depends on the concrete implementations.

4.5 Summary

This chapter presents the capability of representing the clinical process as the leading idea that contextualizes this thesis and shows the main requirements to be covered; some of them considered essential capabilities of such a system. This thesis proposes a system mainly focused over the healthcare professional needs with an active participation by the patient. A system driven by explicit knowledge, which could provide true help to professionals when was needed, and a system semantically interoperable with other systems. It is important to notice that explicitly declared knowledge can be used to argue proposed and taken actions. The fact that the system is declaratively specified by means of ontologies, guarantees technical independence and facilitates the evolution of the system.

Chapter 5

OntoDDB

Ontology based systems need persistence solutions capable of managing ontologies efficiently. At the begin of our work we had chosen RDF/RDFS [20, 42] as ontological language (OWL does not exists at that moment) and Protégé [Protégé], an open source software, as ontology editing tool. Sometime later the system was adapted to OWL. Protégé were able to export ontologies in RDF format using its XML syntax but its capabilities to store models on a database were very limited. The storage models proposed at that moment were very simple and immature for an industrial perspective as stated by Magkanaraki [55], basically consisting of a table where to record triples [resource, property, value]. On the other hand, the interface provided by Protégé with RDF was a file-based approach, not suitable to manage large ontologies or to perform queries on it.

We decided to design a new storage model taking advantage of relational capabilities, making explicit all components defined in the RDFS and OWL specifications.

We stated the following basic requirements for OntoDDB:

- to be able to manage whatever OWL model.
- to achieve a conceptual representation of the OWL model. Our aim is not to store XML documents but OWL models.

- to be easily portable between different DBMS.
- to be very efficient retrieving concepts through a subclass hierarchy.

5.1 OWL-DB

A relational database (OWL-DB) is used for storing ontologies and instantiated data, following an approach similar to Entity-Attribute-Value (EAV) schema. EAV schemas allow changes in the data structure and have proved their utility for clinical applications [4, 91]. The database was designed according to the OWL specification [70]. Based on Theoharis [95], storage schemes can be classified as schema-oblivious (one table is used for storing the statements), schema-aware (one table per class or property is used) and hybrid (one table per meta-class and property instances with different range values is used).

OWL-DB follows basically a hybrid model, which according to Theoharis [95] is the model that achieves the best performance. In OWL-DB there is a table for each OWL meta-class, such as Resource, Class, Property, Domain, Range, etc. The values of property instances are stored in a table according to its range (i.e., resource, string, integer, etc.). An id-based approach is used to identify resources, since the use of shorter identifiers instead of long IRIs results in space savings and performance benefits [29].

An additional single table is used to store all triples defining the ontology. Adding or deleting statements in this table cause triggers to fire and thus the update of the rest of the tables. The statements table serves as interface with other applications. Any application able to manage OWL statements, for example ontology edition tools, can be potentially connected with OWL-DB.

Figures 5.1 to 5.4 are diagrams showing the relational implementation of the main components of the database. Figure 5.1 shows resources defined in a name space, which can be the classes, properties, constraints, literals or statements. Statements are resources too, defined in a model and composed by subject, predicate and object, all of them registered as resources.

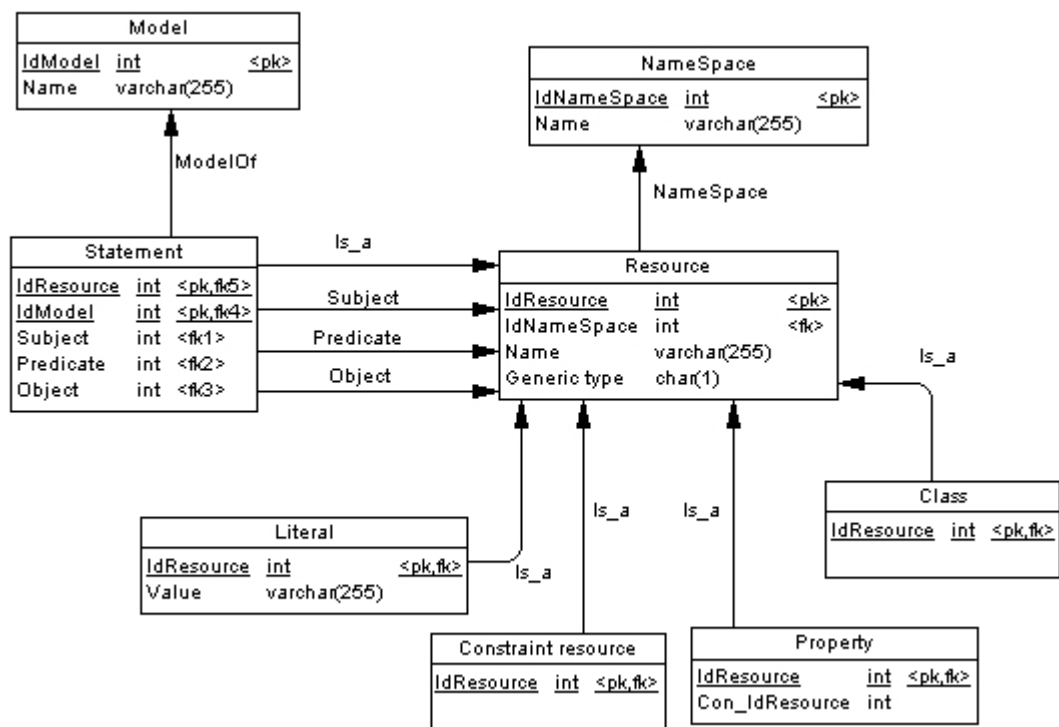


Figure 5.1 - Basic components of OWL-DB

Figure 5.2 shows the definition of the range and domain of a property and the type of a resource. RDF specification allows a resource having several types.

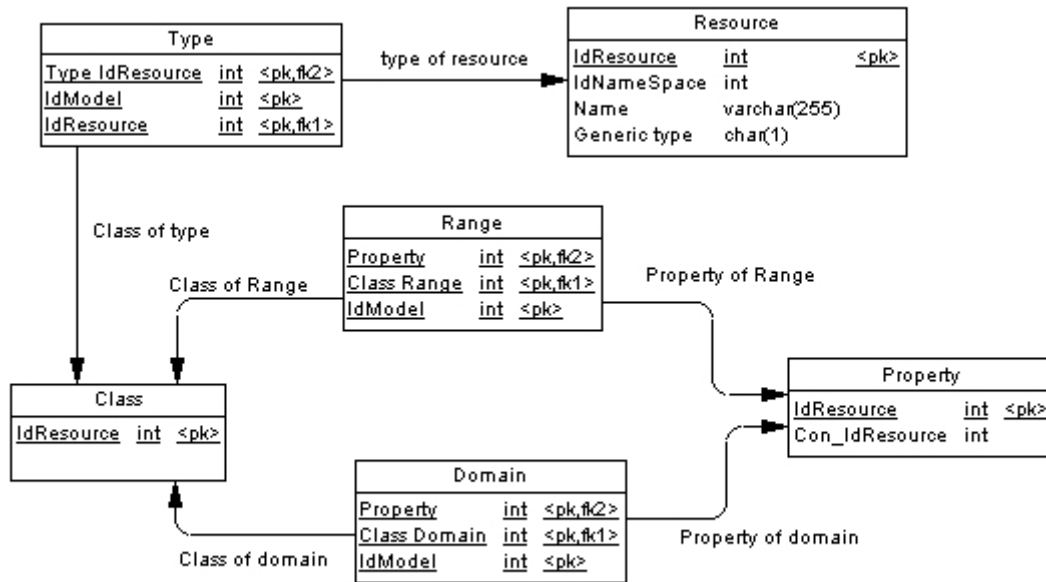


Figure 5.2 - Types and properties in OWL-DB

Figure 5.3 shows the solution adopted to represent languages, a slightly modified version from solution proposed by Motik et al [61]. A literal can be attached to a resource as a label or as a comment. The entity lexical entry represents a specific literal in a specific language, in such a way that the same literal can participate in several lexical entries, each one in a different language.

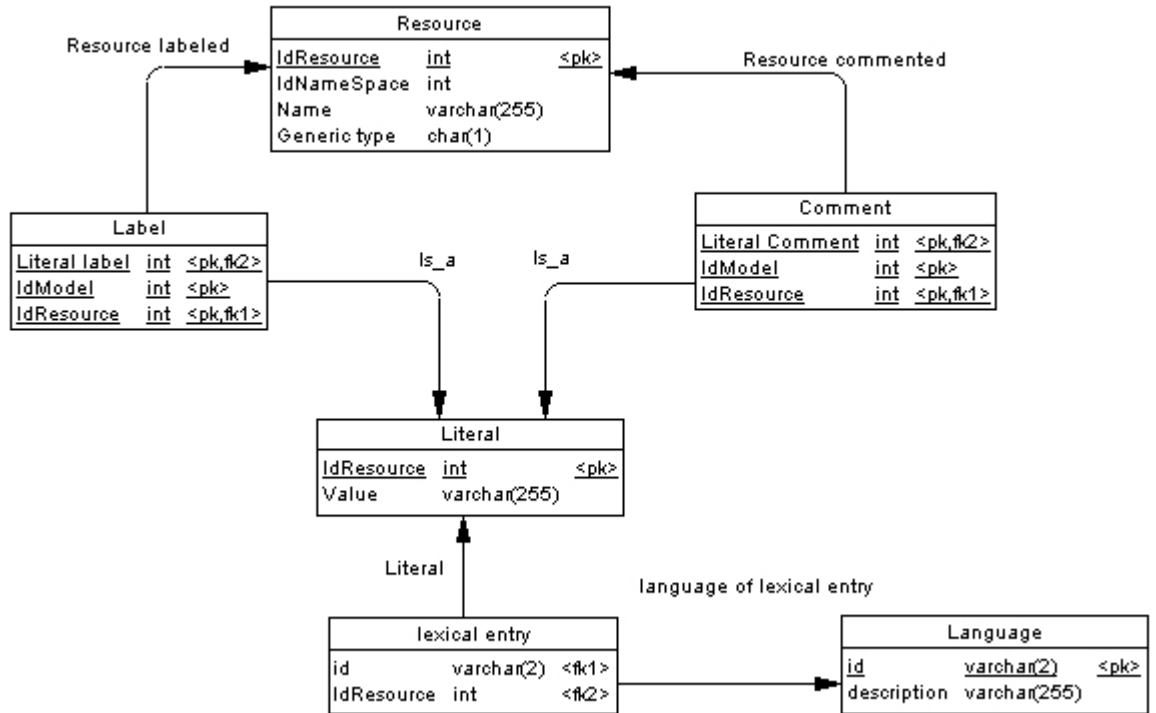


Figure 5.3 - Language representation in OWL-DB

The design of the database is intended for a quick recovery of concepts through hierarchies of classes and subclasses. When only using a statements table, finding the subclasses of a class (through a variable number of levels) is a recursive problem, difficult to solve in the relational environment. To avoid this limitation, subsumption relationships between classes and properties are stored in specific tables, following a nested set model of trees [23]. In this model, each node of the tree is labelled with two numbers (left and right), as shown in figure 5.4.

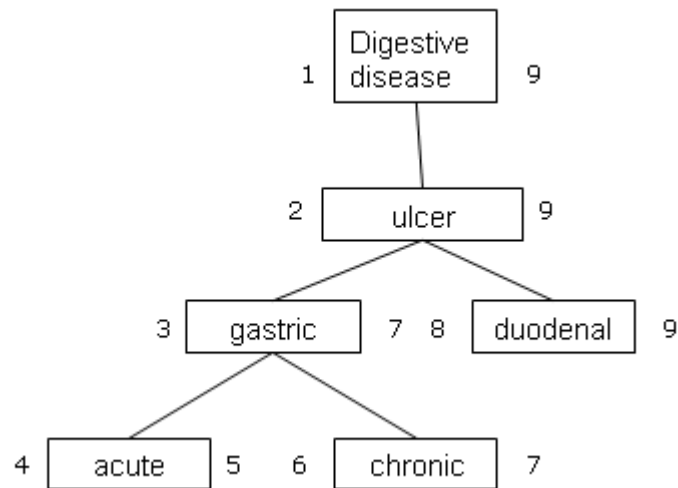


Figure 5.4 - Example of nested set model of trees

Finding all subclasses of a given class A (for example "digestive disease") becomes a very fast process: they are all classes with the right index (or left) comprised between the values of the indices of A. Thus, all concepts defined as subclasses of it, such as "acute gastric ulcer" in the example, will be recovered in a very efficient manner, regardless at what level of depth in the hierarchy they are defined. Nevertheless, this design makes difficult the management of multiple inheritance. Currently, we duplicate the node with multiple inheritance in the class hierarchy, which represents only a small cost in storage space, it remains a unique resource in the database. Figure 5.5 shows the implementation of Class and Property hierarchies.

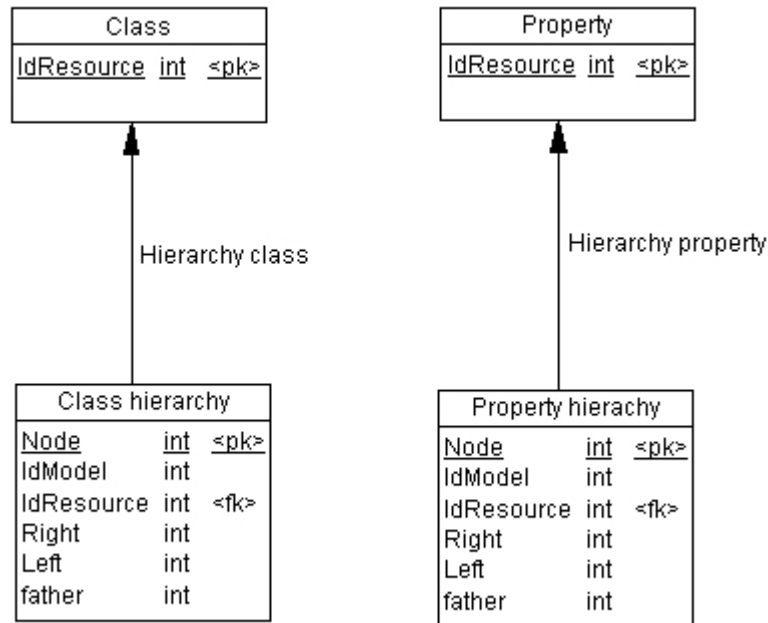


Figure 5.5 - Class and Property hierarchies in OWL-DB

Other applications can interact with OWL-DB using an API built with stored procedures. A set of functions allows retrieving the subclasses, properties and instances of a named class, domain and range of properties, values of instance properties, etc., to extract information from the database.

The system can store in the same database all imported ontologies, maintaining the import relations between the different ontologies.

5.2 OWL-DB Plug-in

The edition of ontologies is based on Protégé [Protégé]. Protégé is a recognized standard for ontology edition, with near 300.000 registered users all around the

world, and able to edit OWL ontologies. The extensibility capability is a very interesting characteristic of Protégé. It is possible to include new functionalities to the tool by adding new plug-ins. A plug-in developed by us, OWL-DB Plug-in, connects Protégé with the OWL-DB module at the storage level. OWL-DB plug-in uses Jena [Jena] to manage OWL statements and to communicate with OWL-DB. This process is shown in figure 5.6.

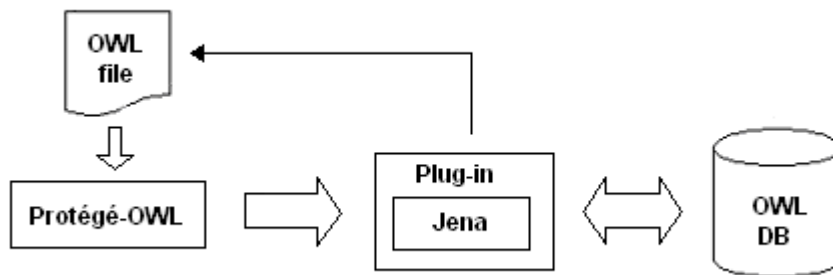


Figure 5.6 - OWL-DB Plug-in

The plug-in implemented is a back-end plug-in. This plug-in consists of a single class, which inherits from the KnowledgeBaseFactory class provided by Protégé. This provides access to some Protégé classes, as KnowledgeBase, allowing class management, properties management and tree interface functionality. Database access is performed using jdbc:odbc. The communication is done by updating the statements table, which triggers the update of the rest of the tables in the database.

Using OWL-DB Plug-in it is possible to load an ontology, previously stored in the database, to be edited in Protégé. The connection parameters to be provided

are DBMS, server IP, database name, username, and ontology namespace. Once a change is done in the ontology with Protégé, the user can choose between saving in the database only the last changes made or replacing the ontology entirely. If the ontology is importing other ontologies, an option is available to save all imported ontologies in the database at the same time.

Using OWL-DB Plug-in, an already existing OWL file in XML format can be uploaded to the database. This is done using the Protégé menu option “Convert Project to Format...” where an option is available to choose the OWL-DB format.

When storing ontologies in OWL-DB from Protégé, a local copy in an OWL file in XML format is automatically generated.

Using this plug-in approach the unique common identifier between the database and Protégé is the resource name. As a user is allowed to modify the name of a resource in Protégé, this could create a conflict. For this reason some changes on Protégé source were needed in order to maintain and manage a list of modified elements.

5.3 OWL-DB OntoLoad

OWL-DB OntoLoad is an application to directly upload an OWL ontology in XML format to the server, feeding the statements table directly instead of uploading it through the editor tool.

5.4 Summary

Real working applications needs solid and scalable storage systems. For building systems using ontologies extensively it is not enough with a file-based approach. We have developed a relational database for OWL, which has been integrated with Protégé.

As result we dispose of a tool that supports the complete process of ontology building and that have all the power and scalability of relational systems.

Chapter 6

OntoCRF

As stated in chapter 2, the lack of available tools in our organization and the disadvantages of traditional database systems, prompted us to seek an alternative to build a platform to deploy efficiently research projects and clinical registries.

The advances in knowledge management tools and methodologies in last years provide the opportunity for a new approach. Ontologies, as explicit conceptualizations of a domain [32], seem well adapted to the task of representing medical data. Since ontologies can be populated with instance data, and deployed as parts of information systems for query answering, ontologies resemble databases from an operational perspective [67]. Languages to represent ontologies, as OWL, have been designed to be extensible and able to accommodate model changes. This flexibility of ontologies is a major advantage of the technology [67]. These characteristics make ontologies suitable to build a conceptual platform on which specific applications can be deployed [47].

In addition, the use of ontologies is more and more common in the healthcare field [16, 34, 40, 86, 88, 90, 94], providing an environment to seamlessly integrate the new information models with existing ontologies.

We have developed OntoCRF (Onto Clinical Research Forms) [52], a framework for the definition, modeling, and instantiation of data repositories. Most important, OntoCRF is capable to face the change at a limited cost, since the implementation of a new repository in OntoCRF does not need database design

or programming. All information required to define a new project is explicitly declared in ontologies, reducing the time and cost of development compared to traditional solutions. The repositories implemented with OntoCRF are accessible via a website for data entry, thus facilitating the collection of distributed data.

The general idea of OntoCRF is to combine the best from two technologies: the expressivity and flexibility of ontologies with the proven robustness and efficiency of relational databases. We have described in the previous chapter a way of using a relational persistence layer to store ontologies [50, 51].

As a general requirement, all information needed for the system to work should be modelled in ontologies. Furthermore, no additional programming should be necessary to implement a new project. To achieve it, each different project has a different ontology that models both the data and the user interface. The ontology indicates which data are needed (i.e. age, sex, etc.) and how to represent them on the screen (a single cell in the first row, a radio button in the second row, etc.) The program code should be the very same for different projects, but being capable of “interpreting” the corresponding ontology to implement different projects.

Although prior work was done with RDF, OWL [70] was finally chosen as modeling language. The justification of using OWL is twofold:

- Be able to reuse existing ontologies to incorporate external knowledge. For example the ontologies stored in BioPortal [66], much of them in OWL format, which are accessible from Protégé.
- Be able to make automatic reasoning in the future. Although not explored yet, we have plans to use reasoners like Pellet [Pellet] for consistency checking, automatic classification, etc.

OWL is a standard with wide support in the Semantic Web community. Thus, tools developed by the Semantic Web community can be directly applied to the data, for example Protégé [Protégé] as ontology editing tool. The election of Protégé is motivated by our previous work on relational support for ontologies [50]. The persistence layer for both models and instantiated data is provided by a relational database.

6.1 Use case presentation

In the following, some examples from current projects will be used to illustrate how the system works. One of them is the registry of the "European Forum on Antiphospholipid Antibodies"; a registry of patients with the "Catastrophic Antiphospholipid Syndrome" (CAPS). This project aims to establish an international data set of all diagnosed patients with the "catastrophic" antiphospholipid syndrome, considered as a "rare disease". For each clinical case the following data are registered:

- Demographic data
- Previous clinical manifestations
- Precipitating factors
- Clinical findings organized by organs
- Laboratory results
- Treatment followed
- Outcome

The data have to be stored in a centralized database to allow periodic statistical analyses on them. In order to allow a decentralized introduction of data, a web based application is needed. A couple of screenshots of the data entry are shown in figures 6.1 and 6.2.

Figure 6.1 shows the list of clinical cases from CAPS registry, and figure 6.2 a concrete case and some of its laboratory results.

The screenshot displays the website interface for the European Forum on Antiphospholipid Antibodies. The main navigation bar includes links for HOME, CONTENTS, GLOBAL RESULTS, Public CAPS Registry (highlighted), LIBRARY, CAPS Registry, and CASES. A search bar is present with a 'Search' button. The central content area is titled 'Clinical case' and contains a table with 14 rows of data. The table columns are Patient Id., Episode Id., Sex, Age, Diagnostic, CAPS as first manifestation, and First manifestation. The data is as follows:

Patient Id.	Episode Id.	Sex	Age	Diagnostic	CAPS as first manifestation	First manifestation
1	1	Female	32	SLE	No	Renal
2	1	Male	43	SLE like	Yes	Pulmonary
3	1	Female	22	SLE like	No	Neurologic
4	1	Female	22	SLE	Yes	Neurologic
5	1	Female	40	PAPS	No	Neurologic
5	2	Female	42	Not available	Not available	Pancreatic
6	1	Female	22	SLE	Yes	Not available
7	1	Female	13	SLE	Yes	Not available
8	1	Female	37	SLE	No	Skin
9	1	Female	33	SLE	No	Not available
10	1	Male	14	SLE like	Yes	Adrenal
11	1	Female	19	PAPS	Yes	Not available
12	1	Female	25	PAPS	No	Bone
13	1	Male	57	PAPS	Yes	Testicular
14	1	Female	11	SLE	Yes	Pulmonary

At the bottom right of the page, there is a link: [Problem? Contact technical support!](#)

Figure 6.1 - The list of clinical cases in CAPS

Figure 6.2 - A clinical case in CAPS

The panel on top left allows the navigation through the different parts of the registry. The windows on the right, which constitute the formularies to fill in, are composed of single cells, combo boxes, check boxes, radio buttons, etc. to introduce and visualize the data.

6.2 General architecture

The general architecture is shown in figure 6.3. OntoCRF is built on top of OntoDDB and is composed of the following modules:

- OWL-DB, a relational database for storing the ontologies and instantiated data. This module is part of OntoDDB.
- An ontology editor based on Protégé and connected with OWL-DB. This module is part of OntoDDB.
- A graphic user interface based on Liferay [Liferay]
- A metamodel describing the primitives of the system.
- An application for data extraction in the back-end.
- An application for ontology upload in the back-end.

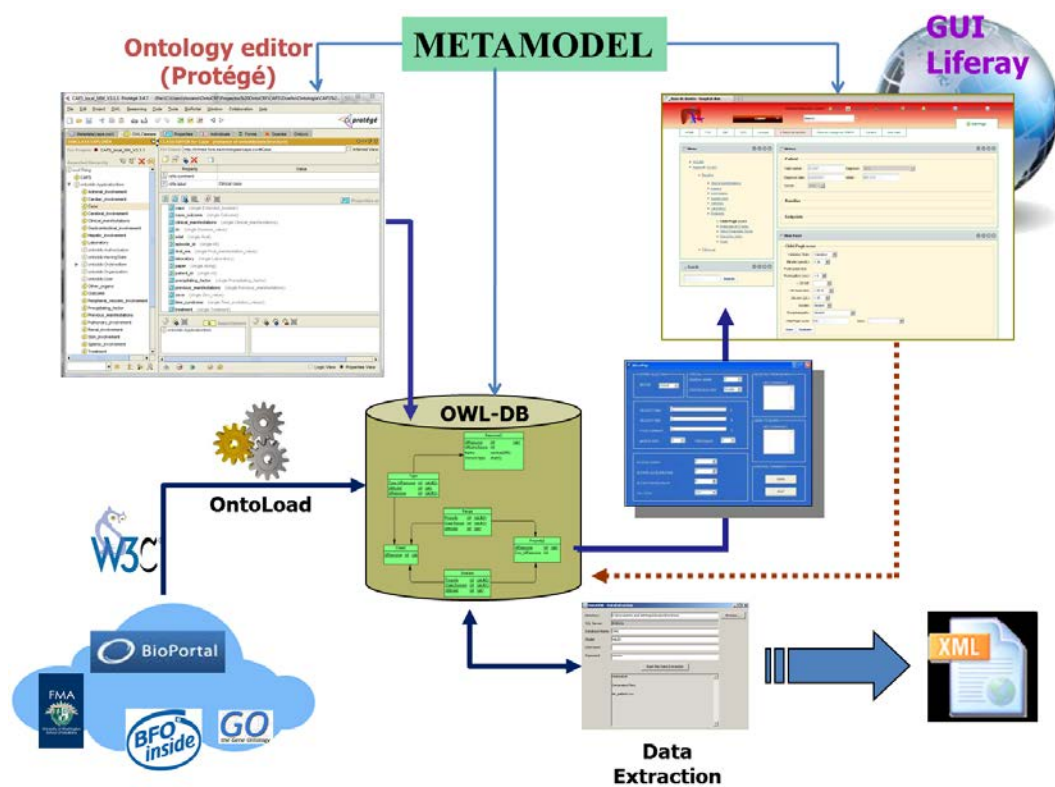


Figure 6.3 - General architecture of OntoCRF

6.3 Ontology authoring

The edition of the ontology is based on OntoDDB. With OntoCRF, the data to be registered are modelled in an ontology. Simplifying, and making a parallelism with relational databases, tables become classes and columns become properties.

Figure 6.4 shows a snapshot of the CAPS ontology. Some classes representing the main groups of data to be registered (Case, Precipitating_factors, Previous_manifestations, Adrenal_involvement, Cardiac_involvement, Laboratory, Treatment, etc.) can be identified.

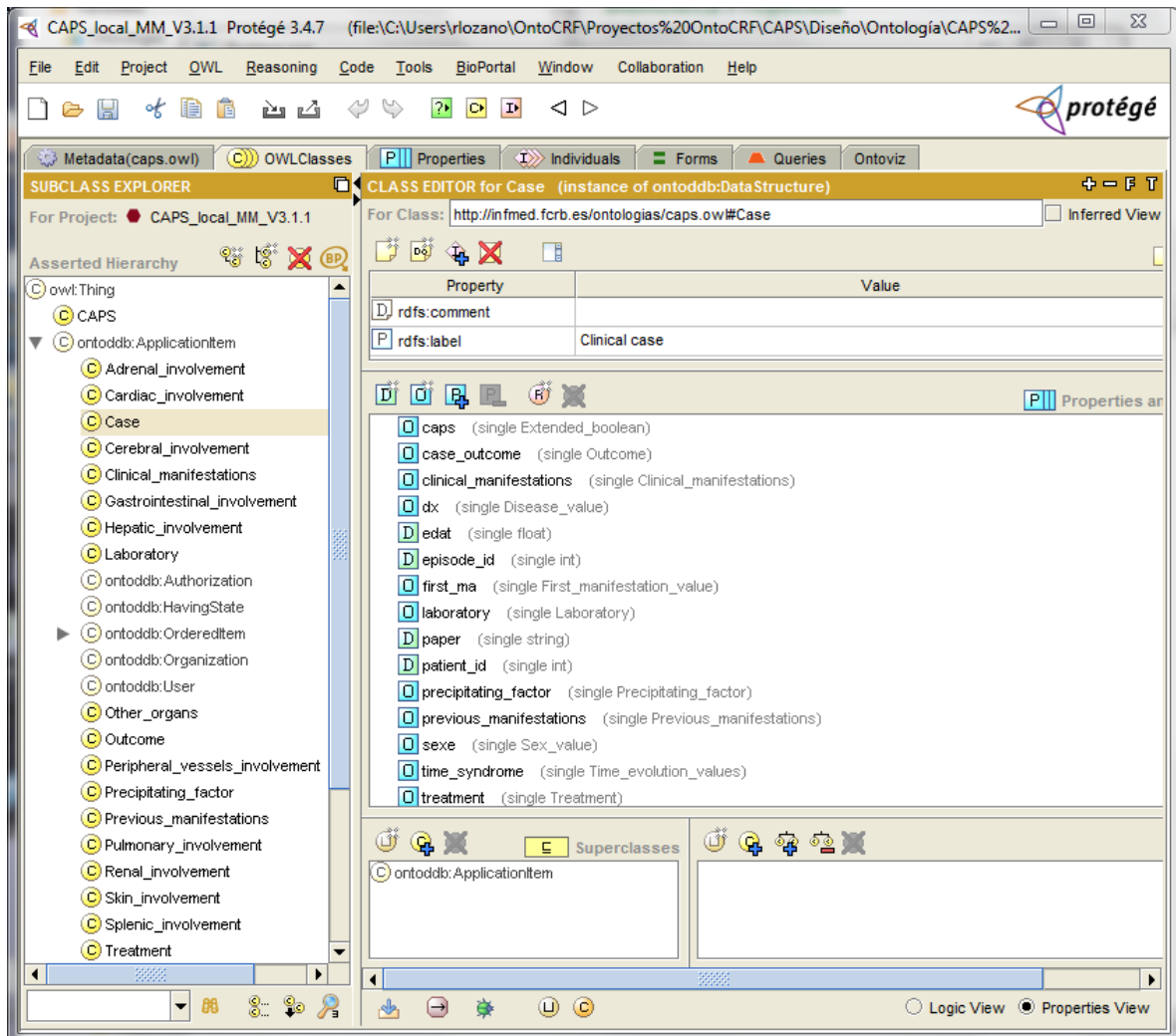


Figure 6.4 - Ontology edition with Protégé

OWL and Protégé support additional functionality, as subclasses, metaclasses, etc. which together with the metamodel allows to use Protégé as a twofold design tool:

- A kind of database design tool to define the data, its structure and properties.
- A graphic interface design tool to define how the data will be presented to the user.

6.4 OntoDDB-MM, The Metamodel

OntoDDB-MM, the OntoCRF metamodel, is an ontology composed of a set of metaclasses, classes and properties that defines the available elements that can be used to build an application. These elements are recognized and used by the portlets to create the GUI. Figure 6.5 shows the main hierarchies.

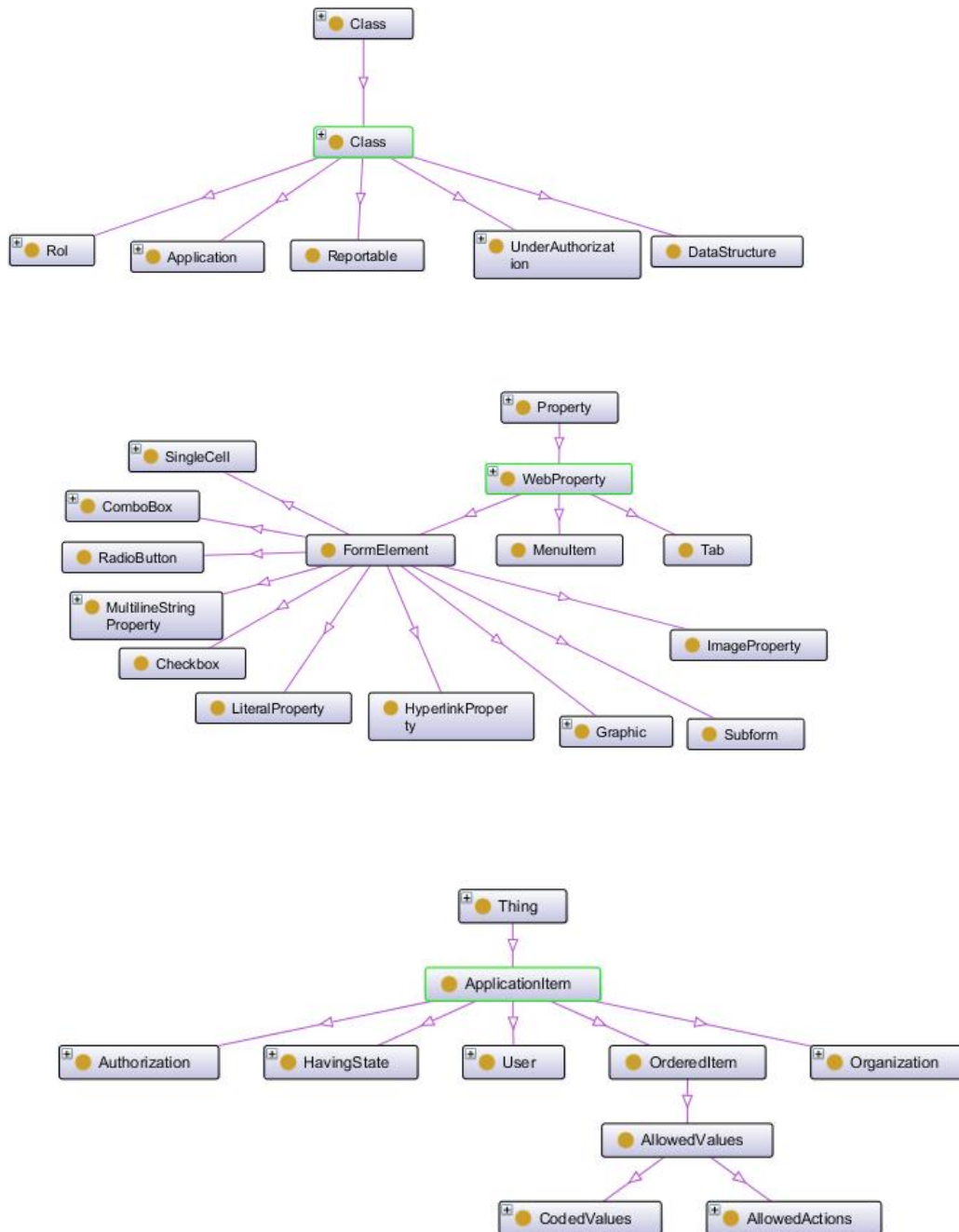


Figure 6.5 - OntoDDB-MM

In this metamodel, an application is represented by an instance of the metaclass *Application*. In the CAPS registry example, the application is represented by the class *CAPS*. The different forms are represented by instances of metaclass *DataStructure*, for example *Case*, *Previous_manifestations*, *Clinical_manifestations*, etc. At the same time, these classes are subclasses of the *ApplicationItem* class.

A *DataStructure* can have several properties, some of them *DatatypeProperties*, some other *ObjectProperties*. In our example, the properties *previous_manifestations*, *precipitating_factors*, *clinical_manifestations*, etc. are instances of both *ObjectProperties* and *MenuItem*. This means, on one hand, that they are properties linking *Case* with other data structures and, on the other hand, that they are menu elements. Figure 6.6 shows an example.

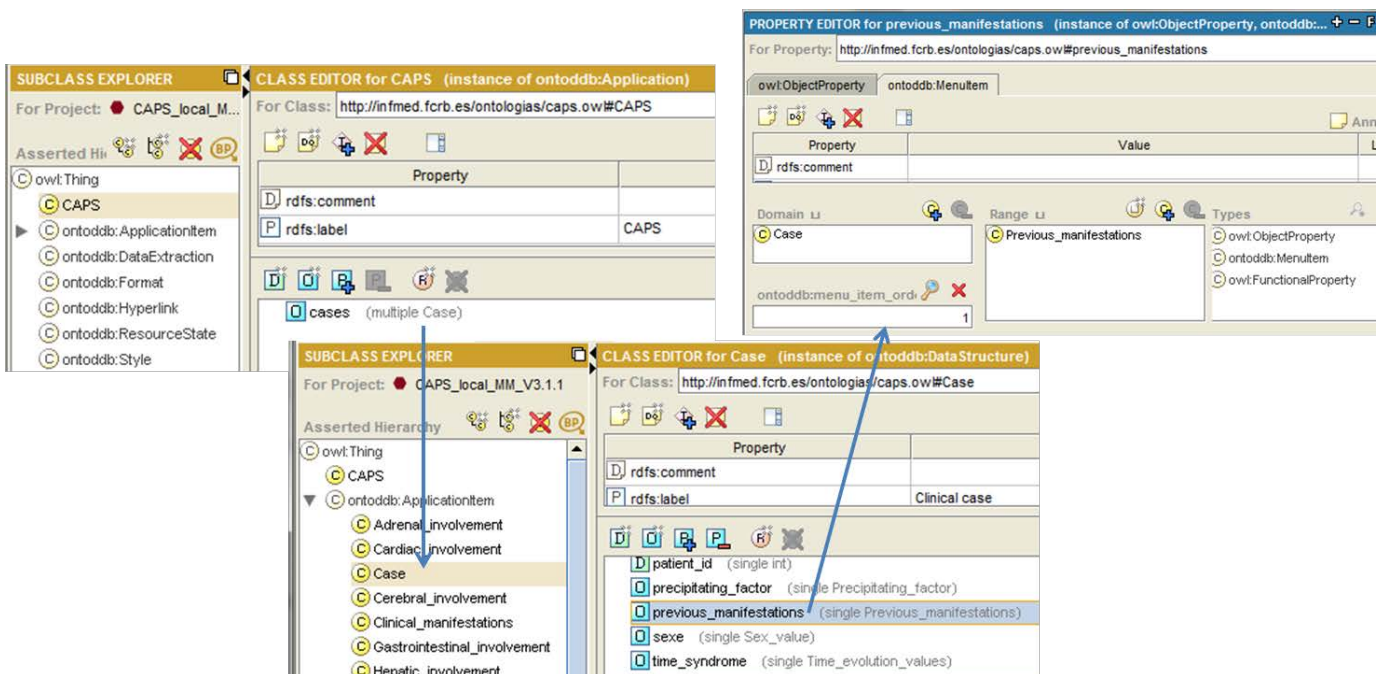


Figure 6.6 - Example of menu elements in OntoCRF

Each form field is an instance of one of the subclasses of *FormElement*, which determines its behavior:

- *Checkbox*.
- *Combobox*.
- *Graphic*.
- *HyperlinkProperty*.
- *ImageProperty*.
- *LiteralProperty*: to represent literals, do not expect a value.
- *MultilineStringProperty*.
- *RadioButton*.
- *SingleCell*:
 - *Password*
- *SubForm*: *not implemented yet*

To manage the form fields, the *FormElement* metaproperty introduces the following facets:

- *webColumn*: the relative column in the form where the field will be shown
- *webRow*: the relative row in the form where the field will be shown.
- *webDescriptionProperty*: a flag to mark fields that are part of the description of the corresponding object and are shown in the headers, list, etc.

- *webMandatoryProperty*: a flag for fields do not allowed to have a null value.
- *webIdProperty*: a flag to mark fields that constitutes the Id of the corresponding data structure. This means that is mandatory to fill in the field and that the value must be unique.
- *webEditionDisabled*: a flag to avoid a field be edited.
- *webDirectlyDependent*: a flag to identify depending objects. The objects that are values of these properties, cannot exists without the object that has this property.

The metamodel allows to indicate that there are constraints on the values to be used which each field. This is done by creating a subclass of the class *AllowedValues* for each field to be constrained. This class is a subclass of *OrderedItem* and their instances can have a relative order between them. If the subclass *CodedValues* is used, instead of the class *AllowedValues*, each of the different options can have an attached code. This mechanism is similar to the method used by Rector et al. to constrain the codes to placeholders [77].

As an additional feature, the system can notify via email the creation of new instances. This is useful to notify adverse events, for example. To do that, the class whose instances have to be notified has to be subclass of the metaclass *Reportable*.

Other classes as *Role*, *UnderAuthorization*, *Organization* and *Authorization* should allow the management of access permission to the different resources, but are only partially implemented at the present moment.

6.5 The Graphic User interface

Using Protégé and OWL-DB is enough to instantiate the ontology in a centralized repository. However, this would not be a suitable interface to an end user.

The user interface is built with portlets based on Spring MVC and deployed in Liferay. The business and controller levels are supported by Spring and the view level by JSP with JSTL. The screen presentation and direct interaction is made with HTML, Javascript and JQuery. With this approach, the end user only needs a web browser to interact with the system.

The GUI is created dynamically: the navigation menu, the components generation, and all objects in general, are created dynamically following the specification of the ontology. The portlets access directly to the OWL-DB stored procedures. Then the information about the application, expressed in the ontology, is used to build the web pages on the fly, as shown in figure 6.7:

Figure 6.7 - Example of form elements in OntoCRF

6.6 Data extraction

Data extraction module aims to allow periodic extractions of stored data to analyse them. This is done by invoking a Java application that will ask the user to provide the connection parameters.

The output of this application is a set of XML files containing the data. These files can be imported to a conventional relational database or a statistical package to be analysed.

Which data has to be extracted is defined in the ontology as instances of the class `DataExtraction`. This class allows the user to specify which class of the application should be extracted and whether the value of their object properties must be traversed recursively or not.

Another available functionality can transform the entire ontology in a relational database. In this case the output is a SQL script. This functionality can be used on a daily basis to maintain a relational version of the data and use existing analytical tools. This approach has been used in some projects. Pentaho [Pentaho] has been connected to the relational version in order to automatically provide some descriptive data and allow queries.

6.7 Limitations

The metamodel of OntoCRF is not capable of process representation, hence not being able at the moment to manage explicit knowledge related to processes. The data extraction capacity is also limited. Nowadays, the final user cannot perform direct consultations over the server. Instead, data need to be previously extracted. Nevertheless, this limitation is currently being addressed and some SPARQL tools are being tested with the aim to be integrated with OntoCRF.

6.8 Summary

OntoCRF is a framework for the definition, modeling, and instantiation of data repositories. It does not need any database design nor programming. All required information to define a new project is explicitly stated in ontologies. Moreover, the user interface is built automatically on the fly as web forms, whereas data are stored in a generic repository. This allows the immediate deployment and population of the database as well as instant availability online of any modification.

Chapter 7

OntoCR

Next objective was to implement semantic interoperability into the system. The choice was to design a clinical repository conforming to CEN/ISO 13606 standard. This repository is able to automatically be populated with external CEN/ISO 13606 extracts. At the same time generates CEN/ISO 13606 extracts with the data it contains. It is out of the scope of the system to incorporate or to generate data using other EHR representations than CEN/ISO 13606. Nevertheless, it would be possible to use different EHR representations following the same approach. This chapter describes OntoCR, a clinical repository driven by ontologies, conforming to CEN/ISO 13606 standard. In this iteration of the repository, we have concentrated on parts 1 and 2 of the standard, leaving aside audit and security issues.

The CEN/ISO 13606 standard does not define an internal architecture or database design of the storage system. Therefore, when implementing this standard over an existing EHR system, a translation layer is needed between the internal structure of the database and the 13606 RM, to transform the data.

One way to avoid complex translations from the internal structure to the Reference Model is to use the latter in the persistence layer. This is the approach followed in OntoCR. Clinical data are stored in the database as instances of archetypes and as 13606 RM constructs: folders, compositions, sections, entries,

clusters and elements (see section 1.7). In this way, adding data from an extract to the repository and creating extracts from the repository are simple processes.

OntoCRF [52] has been described in the previous chapter. In OntoCRF, the repository is independent of content specification. All the required information to define a new project is explicitly stated in ontologies. The user interface is built automatically on the fly as web pages, while data are stored in a generic repository. In OntoCRF, data structures are modeled in an ontology according to a specific meta-model that defines the available elements that can be used to build an application, mainly for GUI definition.

The proposed solution is to build OntoCR [53] by extending OntoCRF, thus achieving a native CEN/ISO 13606 clinical repository driven by ontologies.

The approach followed in OntoCRF is similar to the dual model of CEN/ISO 13606. In OntoCRF, the information model corresponds with both the database model and the meta-model, which remain unchanged. The knowledge model corresponds with ontologies specifying the content. Using Protégé [Protege], the OntoCRF meta-model has been extended by incorporating both 13606 RM and 13606 AM, enabling the capability of representing clinical data that conforms to the CEN/ISO 13606 standard. The overview of the general architecture is illustrated in figure 7.1. Archetypes in OWL 1 format can be uploaded into the system to build specific applications and specific patient data can be queried as CEN/ISO 13606 extracts.

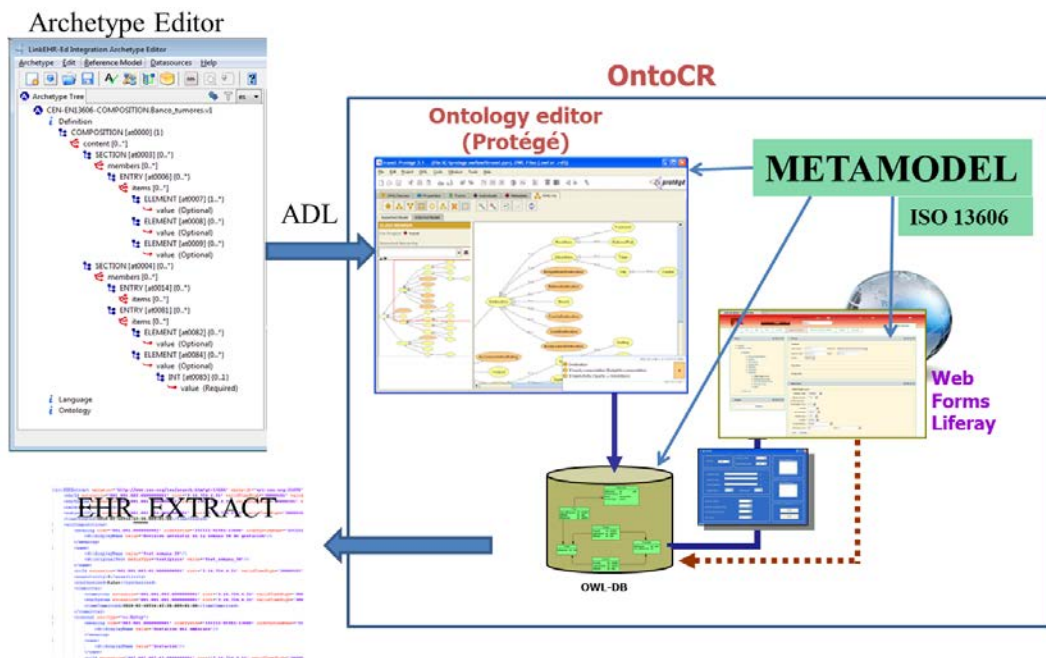


Figure 7.1 – Overview of the general architecture of OntoCR

OntoCR is a native CEN/ISO 13606 system, driven by ontologies, organized in a three-layered model:

1. meta-model layer: meta-model containing definition of CEN/ISO 13606 Reference Model and Archetype model, ISO 21090 data types, SNOMED CT structure, and OntoCRF meta-model for GUI definition;
2. model layer: representation of each detailed archetype and application;
3. instances layer: specific patient data - web forms are automatically generated on the fly to record data specified by archetypes.

7.1 Meta-model description

In OntoCR, we intend that not only communication, but also storage and data capture be available. CEN/ISO 13606 Reference and Archetype Models, ISO 21090 data types, and SNOMED CT structure are modeled and integrated as OWL ontologies, which constitute a meta-model for constructing detailed archetypes.

As a design principle, OWL elements are used when available. This avoids, for example, the need to create a new class list to represent lists of properties, because OWL has its own model to define properties.

Although defined classes in the standards will generally become classes in the meta-model, and defined properties in the standards will generally become properties in the meta-model, the modeling process is not so straightforward. CEN/ISO 13606 and ISO 21090 standards are information models, not conceptual models. To achieve the meta-model, there has to be a transformation process from the information models specified in the standards to the conceptual models represented by ontologies. For example, when a codified item is communicated between two different systems, using the ISO 21090 CD data type, not only the corresponding code is communicated, but also the code system to which it belongs to, the code system version, the code system name, etc. A direct representation of type CD as a class with properties `code`, `codeSystem`, `codeSystemName`, `codeSystemVersion`, and so on, would result in repeatedly storing data about the code system used, with each code, which would be very inefficient. Representing the relationship between codes and code system conceptually, as explained below, produces a more efficient, consistent and scalable system.

7.2 Modeling data types: ISO 21090

The first element to be modeled is the data type system used.

ISO 21090 [ISO 21090] standard provides a set of data type definitions for representing and exchanging basic concepts that are commonly encountered in healthcare environments. The different data types are represented as classes, with a number of generalization / specialization relationships among them. For example, a character string is defined in ISO 21090 as:

```
type ST = class (  
    validTimeLow : characterstring,  
    validTimeHigh : characterstring,  
    controlActRoot : characterstring,  
    controlActExtension : characterstring,  
    nullFlavor : NullFlavor,  
    updateMode : UpdateMode,  
    flavorId : characterstring,  
    value : characterstring,  
    language : characterstring  
    translation : SET(ST.NT)  
)
```

In OntCR each one of the classes defined in ISO 21090 has been translated as an OWL class in the ontology, and the generalization / specialization relationships have been maintained. The exceptions are the types defined on a parameter, such

as Set(T), where a new class is needed for each combination of the basic type and the parameter type.

Each enumeration, for example NullFlavor, is modeled as a class, and the different values of the enumeration are represented as instances of the corresponding class.

7.3 Reference model: CEN/ISO 13606

Reference model CEN/ISO 13606 is conceptually modeled in an ontology that imports the ISO_21090.owl ontology. In general, each one of the classes defined in the standard is modeled as a class in the ontology, for example RECORD_COMPONENT, COMPOSITION, ENTRY, etc. Each one of the properties defined in the standard is modeled as a property in the ontology. In this way, a single extract of a particular patient becomes an instance in the ontology.

We have previously commented on the design principle for using build-in elements of OWL. Following this principle, defined associations in the RM, such as EHR_EXTRACT.all_compositions, COMPOSITION.content, CONTENT.members, ENTRY.items or ITEM.parts, are implicit in the structure of the ontology. For example, we did not model the attribute ENTRY.items as a generic property with the domain ENTRY and range ITEM. Each specific ENTRY (for example “Tumour data”) has specific OWL properties (for example size) whose range is an ISO 21090 datatype (for example INT)

Properties which are defined in a layer are filled in with values from the lower layer. For example, the definition of an archetype may contain the patient’s age as a property, the value of which will be filled with the specific patient age. A

three-layer model implies defining properties in two layers: the model layer (the values are provided in the instances layer) and the meta-model layer (the values are provided in the model layer). Therefore, properties defined in the standards can be divided into these two categories. Here are examples of each:

- model layer: `EHR_EXTRACT.subject_of_care` (identifier of the subject of care from whose EHR the EHR Extract was created) and `RECORD_COMPONENT.rc_id` (the globally-unique identifier by which a node in the EHR hierarchy is referenced). These values depend on each specific instance. These properties are defined as properties of classes.
- Meta-model layer: `RECORD_COMPONENT.archetype_id` (the identifier of the archetype node) and `RECORD_COMPONENT.meaning` (the standardised clinical or administrative concept to which the name attribute has been mapped). These values describe the archetype used. These properties are defined as properties of meta-classes.

Properties of instance data are modeled as usual, and a class `RECORD_COMPONENT` has been defined with properties such as `rc_id`. A set of meta-classes has been defined to represent the properties of archetypes and extracts. For example, the meta-class `RECORD_COMPONENT_def` has the properties `archetype_id` and `meaning` among others. The class `RECORD_COMPONENT` is defined as an instance of `RECORD_COMPONENT_def`. The same method has been followed for the rest of elements of the Reference Model, namely `FOLDER`, `COMPOSITION`, `CONTENT`, `ENTRY`, `SECTION`, `ITEM` and `CLUSTER`. The classes with the corresponding name define the properties of instances and the meta-classes with the suffix “_def” define the properties of archetypes nodes.

A challenging task is to decide which properties belong to each category. The approach followed has been to try to figure out which data will appear with the

same value in each instance (for example, the attributes `ehr_system`, `meaning`) and which will appear with different values for each instance (for example, `subject_of_care`, `rc_id`). The first group is likely to be represented at the meta-class level, and the second at the class level.

The result is `EN_13606_RM.owl` ontology shown in Figure 7.2. As an example, the properties of the `CLUSTER` class are shown separated into model and meta-model layers. Ovals with thick border represent meta-classes, the other represent classes. Arrows with closed head represent the relation `subClassOf`. Rectangles show the properties of `CLUSTER_def` and `CLUSTER`. For the sake of readability, the relations of instantiation have been omitted, but Classes `RECORD_COMPONENT`, `FOLDER`, `COMPOSITION`, `CONTENT`, `ITEM`, `SECTION`, `CLUSTER` and `ENTRY` are instances of the corresponding classes ending in “_def”.

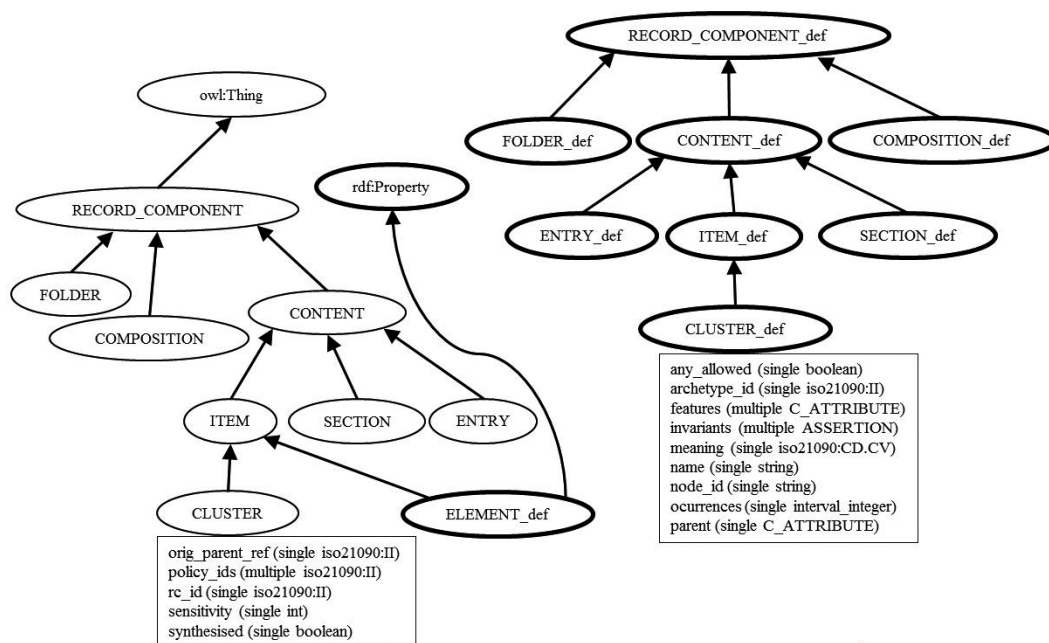


Figure 7.2 - OWL representation of Reference Model.

7.4 Archetype model: CEN/ISO 13606

The process followed with the archetype model was similar to the process followed with the reference model. Associations such as *C_ATTRIBUTE.features* or *C_OBJECT.children* were modeled implicitly in the structure of the ontology as the RDF properties of the class *C_ATTRIBUTE* and class *OBJECT* respectively.

An archetype represents the recording of a clinical concept that will have multiple specific instances, so all properties of 13606 AM are modeled at the meta-model layer as properties of meta-classes. Their values will subsequently be filled in at the model layer when defining detailed archetypes.

Figure 7.3 shows the main elements of the ontology. Ovals represent meta-classes, arrows represent the relation *subClassOf*.

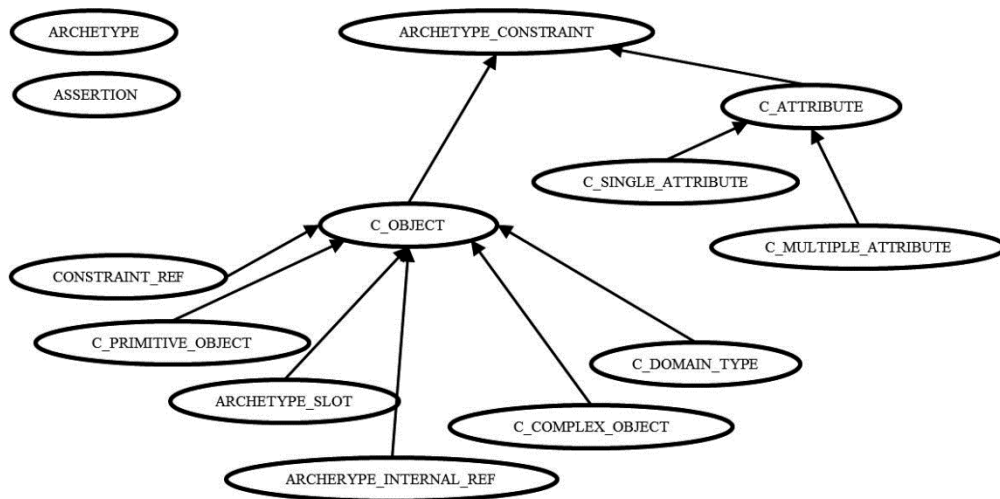


Figure 7.3 - OWL representation of Archetype Model.

7.5 Linking the RM and Archetype model

The link between the RM and AM model is provided in the standard by the property `C_OBJECT.rm_type_name:String[1]`. We have defined the classes of the RM as subclasses of `C_COMPLEX_OBJECT`, as shown in Figure 7.4, where Ovals represent meta-classes and arrows the relation subclass of. By doing so, each instance of `C_COMPLEX_OBJECT`, can be modeled as an instance of the corresponding RM class, thus being a stronger way to link both models.

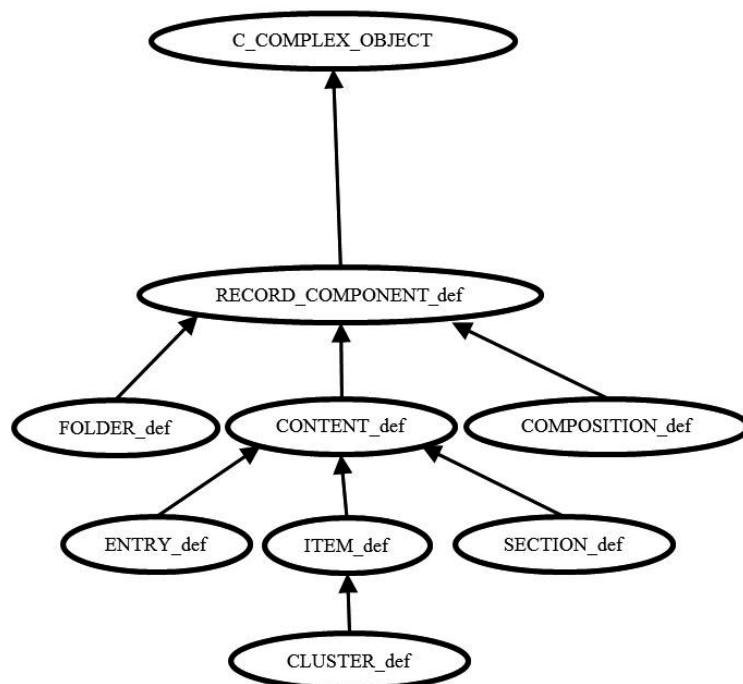


Figure 7.4 - OWL integration between Reference Model and Archetype model.

7.6 Terminologies

CEN/ISO 13606 not only allows for, but also highly recommends the use of controlled vocabularies. ISO 21090 includes data type *CD.CV* to represent coded values. This class includes a set of properties to identify the code system to which the code used belongs, but the values of these properties are strings, so the coding system itself is not conceptually represented.

In OntoCR, each vocabulary is represented by a class which is an instance of the meta-class *Codification_system* and is a subclass of the *CD.CV* class. Figure 7.5 shows an example of SNOMED CT. Ovals with thick border represent meta-classes, the other represent classes. Arrows with closed head represent the relation *subClassOf*, arrow with opened head represent the relation *instantiation*. The class *SCT_CV* is subclass of *iso21090:CD.CV* and *iso21090:CD*, and is an instance of the metaclass *iso21090:Codification_system*. The rectangles represents de properties of classes *iso21090:Codification_system* and *SCT_CV*.

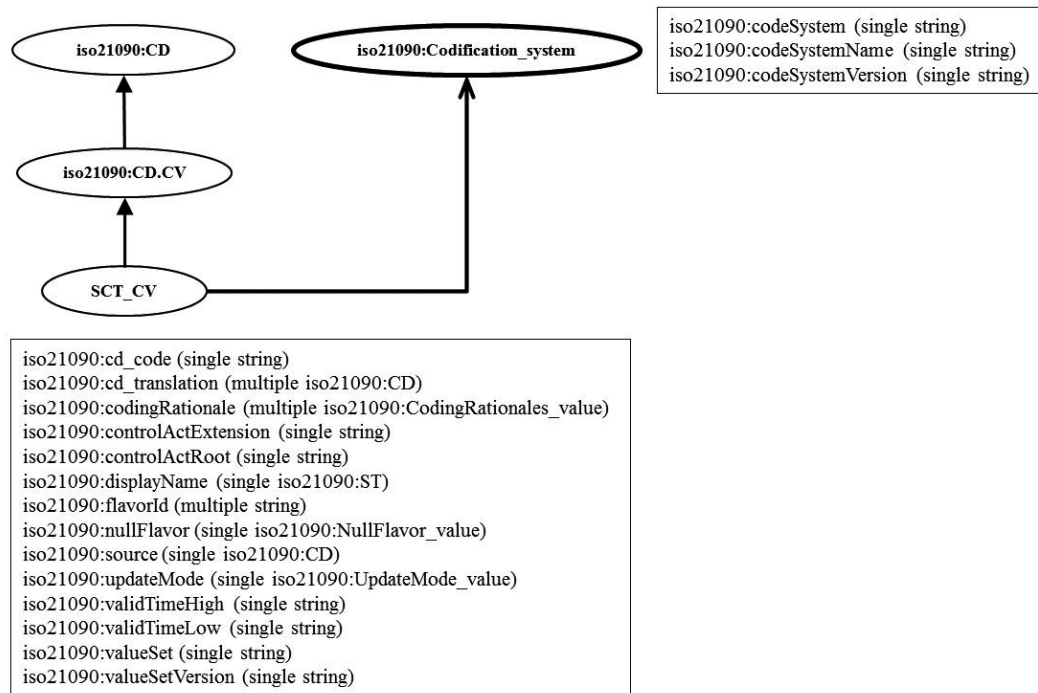


Figure 7.5 - OWL representation of vocabularies.

This approach allows for the non-repetition of common information for each instance. Information belonging to the vocabulary being used is stored only once. For this proof of concept the required SNOMED CT concepts are predefined in the ontology. In a real scenario this model should be integrated with the use of a terminology server and CTS2 services.

7.7 Detailed archetype representation

CEN/ISO 13606 archetypes can be represented as instances of the meta-model described above. They can be created directly, by defining the archetype from

scratch with an ontology editor, or by transforming a previously existing ADL archetype into its OWL version. A software tool has been developed to perform such a translation. It is a Java program that take an ADL file as input and produces an OWL version in accordance with the OntoCR meta-model.

7.8 Summary

OntoCR is a new semantically interoperable clinical repository, based on ontologies, conforming to CEN/ISO 13606 standard. The approach followed is to extend OntoCRF, a framework for the development of clinical repositories based on ontologies, to implement a native CEN/ISO 13606 clinical repository. The meta-model of OntoCRF has been extended by incorporating an OWL model integrating CEN/ISO 13606, ISO 21090 and SNOMED CT structure.

Using a CEN/ISO 13606 based system, an indefinite number of archetypes can be merged (and reused) to build new applications. Our approach, based on the use of ontologies, maintains data storage independent of content specification. With this approach, relational technology can be used for storage, maintaining extensibility capabilities.

Chapter 8

POMR (A model of EHR)

Chapters 5, 6 and 7, describe the technological basis needed to give support to the model proposed in this thesis. The current chapter defines the conceptual model that represents patients' clinical data.

Healthcare is provided through activities in the healthcare and the clinical processes reflecting the interaction between a subject of care and healthcare professionals. As has been indicated in chapter 2, healthcare tends to be provided through different healthcare organizations. Such fragmentation of the delivery of care makes harder the conceptual consolidation of patient's clinical situation. To conform the principles stated in section 4.1.1 (essential capabilities) in this scenario, semantic interoperability is crucial for ensuring continuity of care. In order that a single patient can be attended by different professionals who use different information systems, a common conceptual model must be shared among them, regardless of each one specific implementation. Moreover this approach cannot be focused on administrative issues, which certainly are specific for each organization, but in the patients' health problems. Models and concepts defined in CEN/ISO 13940 standard [ISO 13940], representing all aspects of the content and context of the healthcare services, provide the basis for such semantic interoperability.

In this thesis we propose a Problem Oriented Model Record (POMR) [100], conforming with the system of concepts defined by CEN/ISO 13940 standard [ISO 13940].

8.1 The Problem Oriented Medical Record

The idea of a problem oriented medical record was firstly introduced by Larry Weed [100, 101] in 1968, for developing an electronic health record system (Problem-Oriented Medical Information System, PROMIS), although Weed's vision for medicine goes far beyond software [102]. Weed was aware that physicians were to cope with multiple health problems at a time in a given clinical situation, and had to read the entire medical record and then sort the data in their minds to know all the patient's difficulties and the extent to which each had been analyzed. There is no evidence that this can be done reliably and consistently. Weed proposed as a solution to orient data around each problem. Each medical record should have a complete list of all the patient's problems, including established diagnoses, abnormal physical findings and symptoms. Once such a list has been established all subsequent orders, plans, progress notes and numerical data can be recorded pointing to the problem to which they are specifically related.

For each health problem daily progress notes should be written in the SOAP format:

- S(ubjective): the subjective impressions taken from the patient's complaints, in narrative form.
- O(bjective): includes data observed and measured by the healthcare provider, such as vital signs, laboratory results, etc.

- A(ssessment): diagnoses, differential diagnosis, risk factors, etc.
- P(lan): the activities to be conducted to treat patient' problems or elucidate the diagnoses.

After developing the POMR concept, Weed developed the PROMIS (Problem-Oriented Medical Information System) at the University of Vermont. PROMIS was organized entirely around the POMR concept, and was driven by a large medical knowledge base. The objectives of PROMIS were to improve individual patient care and to document the care provided in manner whereby the outcomes of the applications of medical knowledge can be studied and the "knowledge" itself updated and corrected [101].

When the University of Vermont tried, in the 1990s, to implement a new electronic health record, their biggest challenge was that no system available had the level of functionality and the richness of clinical knowledge contained in the PROMIS system [102].

8.2 The Health Care Model Proposed

Although this model applies to any healthcare organization, (as far as use case for defining it) a hospital will be used as an example because the belief that it encompasses every requirement whatever the organization may be. It is regarded as a single healthcare model that integrates all the professionals involved which will be performing different roles. Since it is a conceptual model, those aspects related to the presentation of information are left aside.

We propose a POMR model conforming with the system of concepts defined by CEN/ISO 13940 standard [ISO 13940]. In the following sections we describe the

main concepts of the model. If not said otherwise, we assume the concepts definitions of CEN/ISO 13940 standard.

A healthcare process could be motivated by any health problem and include any set of activities related to the interaction between a subject of care and healthcare professionals. Both the input and the output of the clinical process are health states, described as observed aspects of each health state. Figure 8.1, taken from CEN/ISO 13940 standard [ISO 13940], shows these concepts.

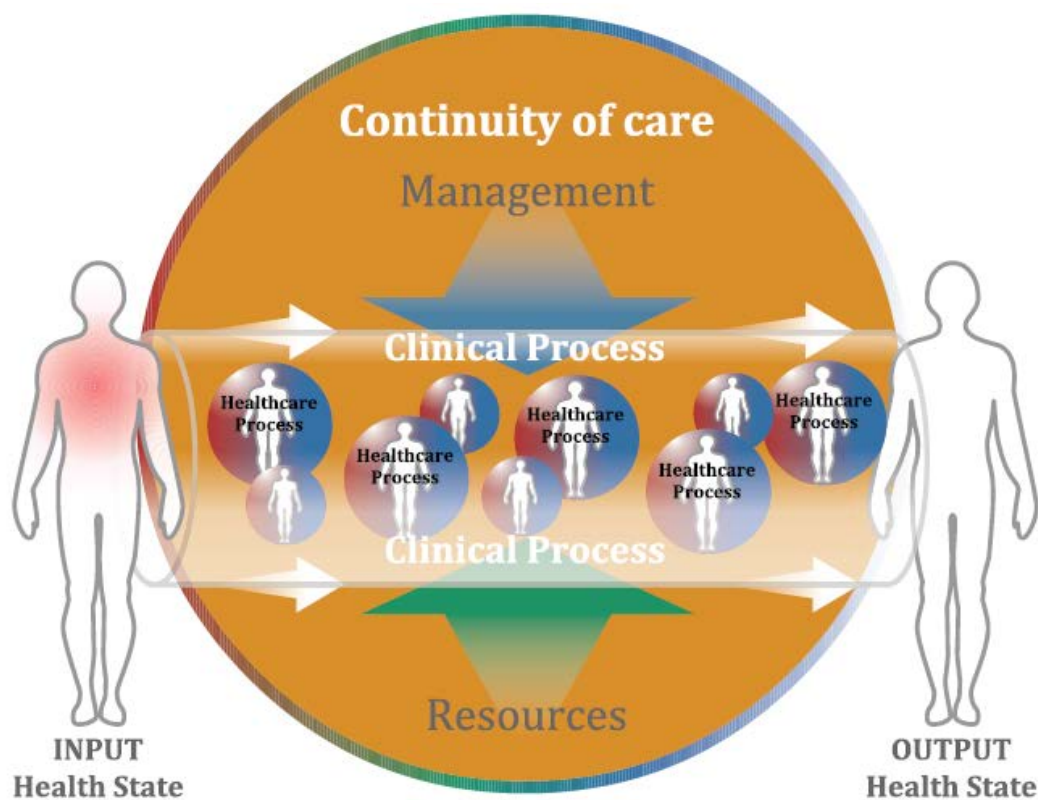


Figure 8.1 – The clinical process in continuity of care. From CEN/ISO 13940 standard [ISO 13940]

8.2.1 The concept of health problem

CEN/ISO 13940 standard [ISO 13940] defines the concept of *health issue* as the "representation of an issue related to the health of a subject of care as identified by one or more healthcare actors"; and consider that a health issue can correspond to a health problem, a disease, an illness or another kind of health condition.

In the model proposed, a health problem is seen as any perception of impaired health or knowledge of health risk factor perceived either by the patient or by a healthcare professional. This concept can be equated to the *health issue* concept defined by the CEN/ISO 13940 standard.

A patient's health problem list must be unique, although two large groups are considered:

- The problem list itself, including diagnosis, signs, symptoms and syndromes.
- Potential problems/Health risk factors including
 - Allergies
 - Carriers
 - Risk factors , including genotype
 - Toxic habits. A dependency problem would be a problem in itself.

Any other health professionals who steps in the process (physician, nurse, social worker) can propose to add new problems to this list.

8.2.1.1 Health problem management

The list of health problems is an essential tool for professionals involved in patient management, so it must be a dynamic entity with all the necessary functionality for supporting the daily healthcare process. Along a patient's care process, the health problems presented and the relationships among them will change as a result of the use of clinical method, and they will be at each point a reflection of the professional thinking process.

In the health problem management three possible actions have been identified:

- **Reclassification:** Involves replacing one problem for the other in light of clinical evidence. This action is carried out when either a health problem or a diagnosis have been considered as such incorrectly, and further data and facts show that what was going on with the patient was due to a different reason. For instance, to reclassify acute pancreatitis as a perforated ulcer. The wrong problem will be replaced for the right one in the list of problems, but the information about the replacement will be kept stored.
 - The new problem must bring it all the information related to the old problem, such as starting date, etc.
 - It should therefore be possible to reclassify a problem whether or not it has components, or if it has elements of any other problem.
 - The replaced problem should not be removed from the system. It must be possible to track its origin, persistence, etc.
 - A problem can be reclassified either by the responsible of it or by the global responsible of the problems list (see section 8.2.4 for a discussion about responsibility over problems).

- **Aggregation:** Several health problems can be grouped when considering that they could be part of a single problem of higher level, even though, at some point, this problem is still unknown. Aggregation doesn't imply a cause-effect relationship between the main problem and the added ones. For example, some syndromics clinical manifestations which could be considered as a whole but without any clear cause.
- **Subordination:** Either a health problem or a problems aggregation can be subordinated to another health problem when considering that those have a causal relationship with an upper level problem. For example, subordinate “abdominal pain” and “ fever” to “Acute pancreatitis” in an individual patient.

Aggregation and subordination actions imply a refining of diagnosis and allow dealing with a number of issues as a whole, although it should also be possible working on each individual problem following its evolution. Furthermore a health problem can be either aggregated or subordinated to several health problems simultaneously.

Aggregation is mainly a day to day work tool that allows establishing possible relationships between health problems. In most cases, at the end of a hospitalization episode aggregated problems would not remain, and the unique kind of relationship between problems will be subordination.

8.2.1.2 Health problem types

Two types or criteria are used to determine the potential states of a health problem.

By its activity over time:

- **Active:** Ongoing health problem that at a certain point in the process requires some action, for example dyspnea, fever, etc.
- **Latent:** Health problem that has been active, that it isn't definitively solved, not needing any other action in the present time that control, but that could be active again in the future. For instance, Hodgkin's disease in remission.
- **Solved:** Health problem that has been active sometime; has now been solved; and either there is not possibility for reactivation or it's a very slight possibility; and requires neither special control nor follow-up. For example acute appendicitis treated with an appendectomy.

By its relevance:

- **Relevant:** A health problem with enough significance that whatever health professional caring for the patient will be aware of it. Therefore should remain visible in the problems list.
- **Incidental:** A health problem that is not considered relevant.

Although it is questionable how to assign relevance to a health problem, it can be considered by definition as relevant any problem either active or latent. Once has reached resolution, the person in charge of the problem should evaluate whether is still relevant or not and accordingly must be reflected in the problems' list. Everything that in the traditional paper based medical records, should be included in the medical history section, must be seen as relevant. Shouldn't be considered as relevant any incidental problem that has appeared and has been solved in a single episode, and will not have effect in the patient's future health

and it isn't important enough as to be taking into consideration for futures episodes.

8.2.2 The work plan

Every problem, either active or latent, can trigger a working plan which may include several actions and in some way should be involved in a clinical guidelines system.

Although any action in the working plan should be motivated by at least one of the items in the patient's list problems, is seen that this relationship doesn't have to be explicitly included each time in the information system, in the interest of improving its functionality.

Actions that could be part of the working plan:

- Tests request. Explicitly must be linked to a problem.
- Consultation/Referral request. May be of a diagnostic or therapeutic nature and explicitly must be linked to a problem.
- Medical orders. Explicitly must be linked to a problem.
- Drug prescription. Every prescription must be explicitly linked to a problem.
- Surgical procedures and any other treatment procedures. Explicitly must be linked to a problem.
- Nursing activities. Any nursing activity is linked either to a problem and/or to risk prevention. It is considered that the information system should not require expressing this relationship explicitly concerning the patients 'general care related activities.

- Other health-care professionals' action planning (social workers, dieticians, occupational therapists...) must be explicitly linked to a problem.

Some of these actions, such as surgical procedures' request, must be validated by Staff members.

Although it's unlikely that a single state schema could be identified for all the actions included in the working plan, there are several possible states for any action that can be considered general enough and therefore widely applicable.

- Requested action
- Scheduled action
- Ongoing action
- Accomplished action
- Documented/Informed
- Cancelled action

8.2.3 The workflow

Figure 8.2 shows the care of patients' working flow within the institution itself, modified from CEN/ISO 13940 standard [ISO 13940].

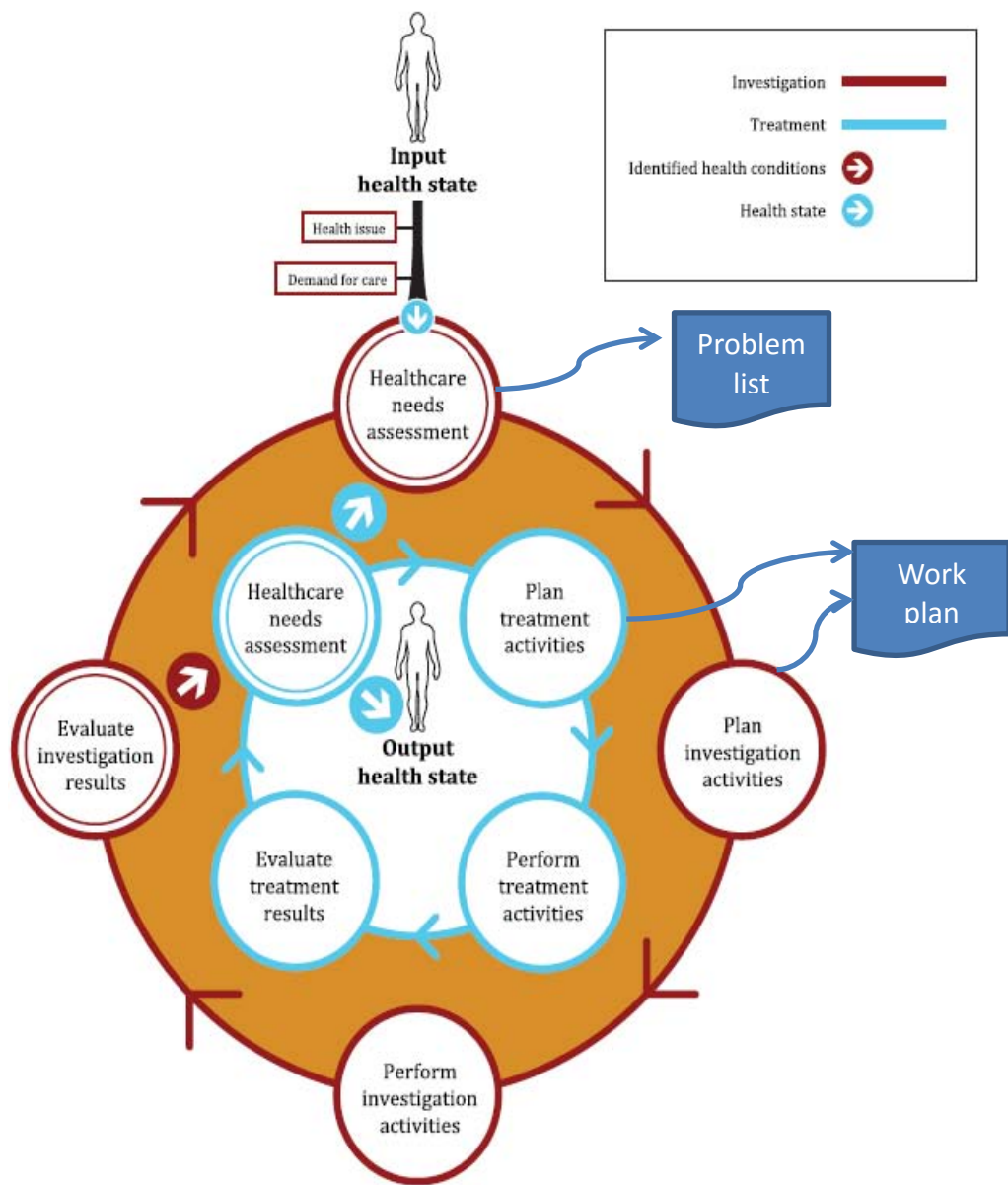


Figure 8.2 – The care of patients’ workflow. Modified from CEN/ISO 13940 standard [ISO 13940]

A patient requires health care for a health problem that it’s the reason for encounter.

The first step is the patient's assessment by the physician using both the information told by the patient and data provided either by him/herself or existing data within the system. As a result of this valuation either the problems list is generated or the existing one is updated.

The next step is to draw up a working plan that includes all those actions deemed necessary for solving problems and their implementation.

From the execution of the actions in the working plan comes a new knowledge (as in additional test, for instance) as well as the amendment of the patient's problems original condition, therefore it is necessary to re-evaluate the patient's situation. This new assessment involves updating the state of the patient's health problems and the problems list. If unresolved problems persist it is necessary to modify appropriately and execute the working plan. This cycle repeats while there are unresolved problems.

8.2.4 Actors and roles

CEN/ISO 13940 standard [ISO 13940] defines the *healthcare actor* as “*organization or person participating in healthcare*”. So, every healthcare professional in an institution is a healthcare actor bearing varying degrees of responsibility in the patients' problems management.

8.2.4.1 Health problem responsible

Irrespective of who has been the particular recorder of a health problem, this issue must always have a single responsible assigned. The responsibility lies at all times either with a department or a functional unit, and in addition any health professional that belongs to them, so any of its members can carry out the tasks:

problems' revision, state's amendments, etc. Can be responsible of health problems physicians, nurses, and other healthcare professionals (social workers, dieticians, occupational therapists). Each health problem may have a different responsible. Therefore the department where the patient it's staying will be responsible for some of them, others are the responsibility of the nursing staff, and a few others are a task of the required Unit. For that purpose the nursing staff will be considered as an only group.

A professional attached to a department/unit we'll have the possibility to transfer the responsibility for a health problem to another professional attached to a different department/unit. To carry out this change is essential that the latter accepts the transfer. While a transfer request is ongoing the health problem responsibility remains unchanged.

The responsibility for a health problem can be transferred to a different institution, for instance from a hospital to a primary care center.

8.2.4.2 Health problem list responsible

In the hospitalization episodes there is a single responsible for the list of health problems management. Usually this responsibility rests on the head of the unit/ward where the patient is hospitalized. In special circumstances may be accepted that the head of a different unit, acting as consultant, may be responsible. This happens when the patient develops during his stay a worst and lasting problem than the one which caused his hospitalization.

In the outpatient clinic the responsibility for the overall list is shared, not being possible to identify a single responsible. Each unit/department physician will be responsible for managing their issues, regardless of the other problems management.

8.2.4.3 Work plan responsible

Basically follows the same schema as in the list of health problems:

- The unit/department, responsible for the patient, and secondly a professional from it, is also considered in charge of the overall work plan management.
- Each particular work plan (nursing, social assistance, etc...) rely on the relevant professional group (nurses, social workers, etc...) as responsible. This doesn't preclude that for each action taken the name of the individual involved is recorded.
- Each work plan action rely on a professional group who is responsible in the same terms as those mentioned in the previous paragraph.

8.2.5 Clinical work stations

The concept of the health problem as the lynchpin on which to organize the patient care implies to transfer adequate operation capacity to the clinical work stations to manage the workday. Although for every patient a single list of problems is considered, the model provides enough elements (problem initiator, responsible, problem state), to allow proper customization of information to show according to each interlocutor at one point.

The system should allow to every professional to determine for themselves which problems are seen as more relevant at each point in time and thereby establish customized lists.

Actions as reclassification, aggregation and subordination are key to problems management and should be regarded as core actions for any implementation.

8.3 Summary

The health care model is the final component of OntoEHR, which constitutes a common conceptual model that represents patients' clinical data, assuring continuity of care. We propose a Problem Oriented Model Record (POMR) [100], conforming with the system of concepts defined by CEN/ISO 13940 standard [ISO 13940]. In this model the concept of the health problem is the axis on which to organize the patient care. Making the health problems explicit and essential objects for clinical management, information systems can be used to support their management. Any action executed on the patient should be motivated by at least one of the items in the patient's problems list. The model proposed assumes the clinical process as defined by CEN/ISO 13940 standard [ISO 13940]. The first step is the patient's assessment, followed by the planning and execution of investigation and treatment activities, and a new assessment of the results obtained. This cycle continues while there are unresolved problems.

Chapter 9

Evaluation

In this chapter we evaluate the specific hypotheses defined in chapter 3. Therefore, we test each technical component of OntoEHR: OntoDDB, OntoCRF and OntoCR, and thus the technical infrastructure of OntoEHR. Although this thesis proposes a problem oriented medical record as paradigm to manage clinical information, is out of the scope to evaluate its feasibility due to organizational implications.

9.1 OntoDDB

As seen in chapter 2, the hypotheses related to the use of ontologies as operative databases were:

H1. A relational database designed following the OWL model is suitable for ontology storage and edition.

H2. Ontologies can be used to semantically integrate specific clinical data.

9.1.1 Repository evaluation

We have used OntoDDB in several projects involving the use of OWL-DB as ontology repository. This represents more than 20 different ontologies about very different domains. In all cases the designed ontologies have been successfully stored in OWL-DB.

To evaluate the adequacy of the system we tested its behaviour to manage a set of different ontologies. The goal was to determine if the time needed to store and retrieve an entire ontology was appropriate. We supposed the size of the ontology to be an important variable, but in a system as OntoDDB was very important to check how the structure of the ontology affects the performance of the system. So, we selected a representative set of ontologies and measured the time to upload and download each one of them.

Input data

We choose a set of seven different ontologies among our running projects, varying in its overall size, number of classes, number of properties and number of instances. Table 9.1 shows the characteristics of the ontologies used in the experiment.

Set up

The upload and download was made between Protégé 3.5 and OWL-DB, and refers to the entire ontology. To measure upload and download times without be influenced by traffic in the net, a local server was used with the following characteristics:

- Operative system: "SUSE Linux Enterprise Server 11 (x86_64)"
GNU/Linux 2.6.32.43-0.4-default x86_64
- CPU: 1 x Intel(R) Xeon(R) CPU E5-4640 0 @ 2.40GHz
- RAM: 4Gb

Analysis

Table 9.1 shows the obtained results

Ontology	Number of Statements	Disc space	Number of Classes	Number of Properties	Number of Instances	Upload time	Download time
1	102.371	48Mb	147	365	26.414	65 sec	12 sec
2	103.652	72Mb	91	288	8.794	58 sec	24 sec
3	191.487	90Mb	81	171	26.408	112 sec	24 sec
4	200.509	98Mb	15.982	90	15.595	311 sec	80 sec
5	264.636	126Mb	258	632	32.317	200 sec	27 sec
6	131.926	74Mb	145	623	7.553	75 sec	17 sec
7	79.350	47Mb	125	535	5.378	44 sec	9 sec

Table 9.1- Results of OntoDDB performance evaluation

Regarding the performance of the system, its behaviour is quite linear. As table 9.1 shows, any one of the variables considered has a preponderant influence. In

general, the upload and download times are proportional to the number of statements. The greater complexity of some ontologies, expressed by a higher proportion of classes and properties in relation to the number of instances, involves a slight penalty. The project number 4, with two orders of magnitude more in number of classes, shows a worse performance, but less than 4 times worse than other projects with similar number of statements. This is owing to the cost of maintaining the class hierarchy tree in the database, mainly when uploading the ontology. In previous versions of the system, each time a class was inserted in the database; all the indexes of the class hierarchy were recalculated. Project 4 showed the lack of scalability and efficiency of this approach, the system was not able to recalculate the indexes and remained working without end. In the current version, the entire class hierarchy is calculated only once, after all classes have been inserted into the appropriate table. This approach represents only a gain of 2-3 seconds for the rest of the projects (not showed in the table), but a radical change for projects with a large number of classes. With this approach, the cost of maintaining the class hierarchy is assumable and, in return, the retrieval of instances at whatever level is trivial.

The above discussion is about uploading and downloading the entire ontology, a task which is performed during the development phase of a project. The user interaction with the system, adding and retrieving data, is no different than with other systems. The user interaction only involves a small set of data, not the entire ontology.

9.1.2 Semantic integration evaluation

According with Stroetmann et al. [92], the primary goal of ontologies and terminologies for interoperability is to enable the faithful exchange of meaning

between machines and between machines and people. To achieve this goal, specific clinical data about patients, coming from different sources, have to be semantically integrated and linked with these terminological resources. We evaluated the capability of OWL-DB to integrate specific clinical data and to perform semantic search of its content in the following projects.

9.1.2.1 SCOPE Project

The objectives of the European SCOPE project [SCOPE] were to remove the impediments to the development of on-line content provision industry in Europe. One of their goals was providing structured knowledge about contents conforming an ontology of the domain so that users can search and retrieve information with a higher level of abstraction than with actual keyword-based systems, which adds value to an on-line.

The content handled by the SCOPE project was the scientific information contained in G&H [G&H], a publication relating to Gastroenterology and Hepatology, and directed to Medical Consultants and General Practitioners. It was anticipated that this content would constitute an ever-expanding corpus of resources, and the need for a sophisticated search facility was clear. Related functionality, the central idea is to implement semantic searching of contents through a knowledge representation system with multilingual support.

More than use metadata to describe the content of a resource as used by Boulos et al [19] we aimed to express the content itself (in a rough way for the moment) in a computable way. The contents of the articles were represented in an ontology that supports query facilities, allowing final users to perform a semantic search of information.

We choose the Unified Medical Language System (UMLS) [UMLS], a long term

project of the National Library of Medicine (NLM) of USA, as terminology reference. UMLS was widely used in similar projects [1, 19]. UMLS has three components, the Metathesaurus, the Semantic Network and the Lexicon. We only use the first one in a twofold way: a) the Metathesaurus contained at that time more than 700.000 concepts and represent a good source to pick up the concepts to feed the ontology; b) as a large number of concepts are represented in several languages, we can build a language independent ontology and use the Metathesaurus as a kind of translator.

Integrating the components

The idea was to build an ontology mixing UMLS concepts and articles. A plug-in which integrates Protégé with the UMLS database was developed. When building the ontology, this plug-in allows searching concepts in UMLS and adding some of them to the ontology under construction. Some information related the concept is added automatically, for example the UMLS code, UMLS semantic code and UMLS semantic description.

Another plug-in allows the integration between Protégé and the database containing the articles to include in the ontology. The ontology contains the link between the contents and the identifiers of the articles. In fact, articles are the leaves of the ontology tree. In this way, a search does not imply checking every article but browsing the tree to reach related articles. Asking the ontology about some concepts will result in a structured list of articles on this subject.

The electronic version of the publication has a web page where a user can search for content. The sequence of facts is as follows:

1. The user submits some words.

2. These words are used to retrieve related concepts in UMLS using appropriate language tables.
3. Concepts retrieved in the previous step are searched in the ontology by its UMLS code.
4. A sub-tree of the ontology, with concepts retrieved, related concepts (subclasses) and articles related to the concepts, is returned to the web page.

This sequence is showed in Figure 9.1

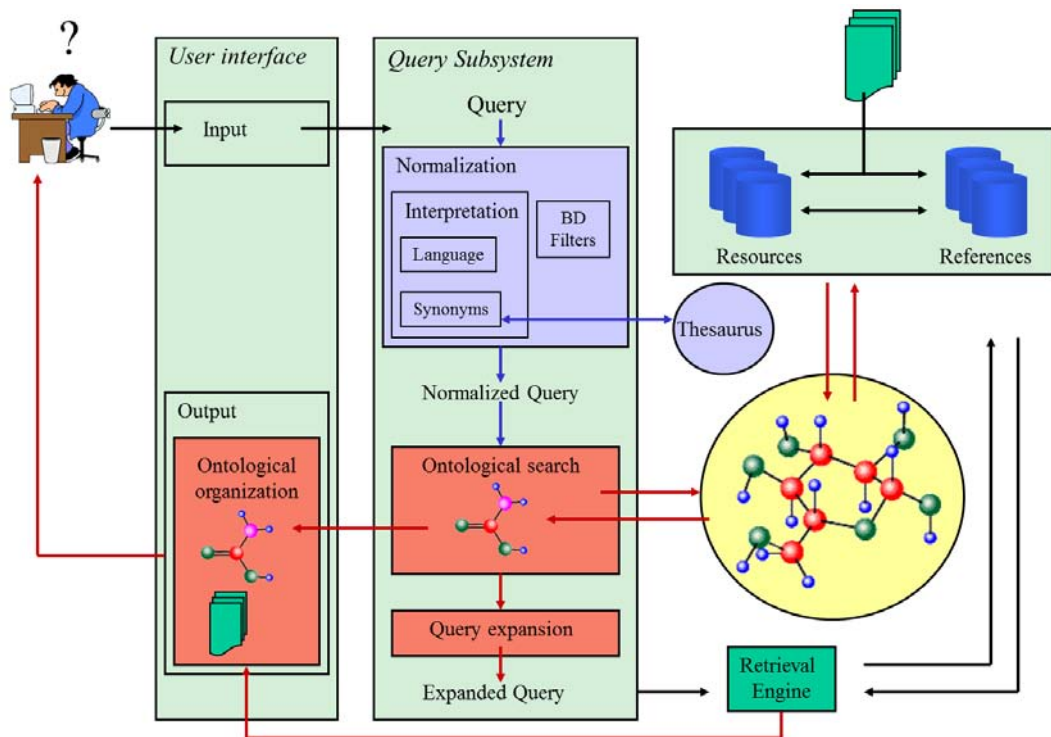


Figure 9.1 – Searching content in SCOPE

Results

Almost all concepts needed in the ontology were found in UMLS. Only 14 of about 300 concepts added to the ontology were not found in UMLS as concepts, but most of them are defined in UMLS as Semantic types, demonstrating the value of UMLS as source of concepts.

The system can perform a search of concepts related with a set of words with an almost immediately answer. From the list of concepts retrieved is very easy to pick up some of them and put them in the desired place at the ontology. New concepts, not present at UMLS, can be created too.

About ontology storage, the tool supports the functionalities for storing different ontologies on a unique database. An ontology can be loaded from the database, modified and saved then back. The system allows a very fast retrieval of articles related with some specific concepts, browsing through the hierarchy.

The result of a query in G&H is a structured tree of concepts related with the words introduced by the user and, for every concept, the available articles are showed. Selecting an article, the user can go to the corresponding web page.

9.1.2.2 INBIOMED Project

The general objective of the INBIOMED project [39] was to build a knowledge based system capable of storing and processing information at several levels (molecular, cellular, tissue, disease, patient) coming from different sources. One of the specific objectives of the project was to integrate clinical data coming from two different hospitals.

The use case

As use case we chose the integration of the CMBD (Conjunto Mínimo Básico de Datos in Spanish) from two hospitals: the Hospital Clínic of Barcelona and the Hospital Virgen de las Nieves of Granada. The CMBD [CMBD] is a registry which contains the personal and clinical information of patients admitted to a hospital and is mandated to be sent to healthcare authorities by any hospital in Spain. Some of the data contained in the CMBD are coded using a common system in all Spanish territory. For example, diagnoses are coded using the International Classification of Diseases version 9 Clinical Modifications (ICD 9 CM) [ICD 9 CM]. Nevertheless, some other data, for example the sex of the patient, are coded using local classifications systems. Not only is the content coded using different classifications, but the name of the variables is different in each case. For example, the variable for birth date is “d_naix” in the case of the Catalan CMBD and “fecnac” in the case of the Andalusian CMBD.

Methodology

In this project each one of the CMBDs was stored in its own database. An ontology was built integrating the definition of both CMBDs using UMLS as conceptual reference. All concepts were taken from UMLS and represented as OWL classes. The codes from each CMBD were modelled as subclasses of the corresponding UMLS concept. Multilingualism was achieved using different labels in English, Spanish and Catalan.

A tool was developed to query the ontology, translate the query to specific queries over each CMBD database, and show the obtained results.

Results

Figure 9.2 shows an example of building a query about female patients with bladder paraganglioma, a kind of bladder tumour.

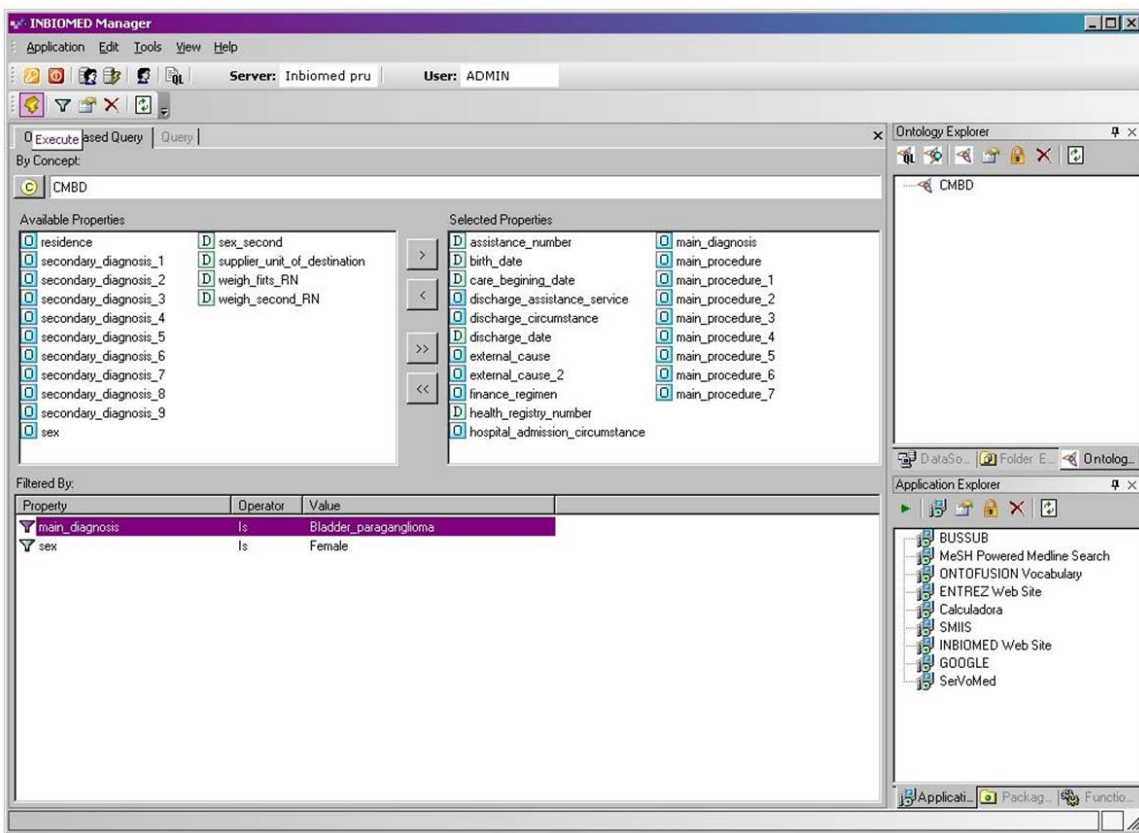


Figure 9.2 – Example of query using the ontology

The tool allows selecting the desired properties for set of available properties and specifying the conditions to match; in this case the main diagnosis to be a bladder paraganglioma and the sex to be female.

The query shown in figure 9.2 is translated to the following SQL code:

```
set @r1=query CMBD_SCS {  
  
SELECT c_ingres as hospital_admission_circumstance, num_assis as  
assistance_number, d_naix as birth_date, d_ingres as care_begining_date,  
ser_alta as discharge_assistance_service, c_alta as discharge_circumstance,  
d_alta as discharge_date, ce2 as external_cause_2, ce1 as external_cause,  
finan as finance_regimen, historia as health_registry_number, c_ingres as  
hospital_admission_circumstance_2, dp as main_diagnosis, pp as  
main_procedure  
  
FROM CMBD_SCS  
  
WHERE (dp='223.3' OR dp='188.9') AND (sexe='1') };  
  
set @r2=query CMBD_SAS {  
  
SELECT tiping as hospital_admission_circumstance, assitenc as  
assistance_number, fecnac as birth_date, servalt as  
discharge_assistance_service, tipalt as discharge_circumstance, ce2 as  
external_cause, ce as external_cause_2, historia as health_registry_number,  
c1 as main_diagnosis, p1 as main_procedure  
  
FROM CMBD_SAS  
  
WHERE (c1='223.3' OR c1='188.9') AND (sexo='2') };  
  
return UNION(@r1, @r2);
```

As can be seen, using the information of the ontology the query link each specific variable name (d_naix, fecnac) to common alias (birth_date) and use the appropriate codes in the WHERE clauses (sexe = 1 and sexo = 2).

A query is performed over each CMBD database and the results are mixed as shown in figure 9.3.

hospital_admin	assistance_n	birth_date	care_begining	discharge_as	discharge_ci	discharge_dat	external_caus	external_caus	finance_res
1	47399799	15/08/1922	18/03/2003	FEP	1	26/03/2003	(null)	(null)	01
2	44787660	04/02/1958	04/02/2003	END	1	10/02/2003	(null)	(null)	01
1	49472389	18/09/1972	22/03/2003	CAR	1	31/03/2003	(null)	(null)	01
1	44928176	07/04/1915	21/01/2003	URM	2	22/01/2003	(null)	(null)	01
2	47399424	22/06/1985	17/03/2003	COT	1	19/03/2003	(null)	(null)	01
1	44457855	21/03/1916	25/01/2003	MDI	1	30/01/2003	(null)	(null)	01
2	44490286	13/01/1926	15/01/2003	OFT	1	16/01/2003	(null)	(null)	01
1	44480100	08/02/1972	13/01/2003	ORL	1	15/01/2003	(null)	(null)	01
1	44928168	07/04/1915	24/01/2003	CIR	6	24/01/2003	(null)	(null)	01
1	47399212	17/06/1940	17/03/2003	COT	1	27/03/2003	E8490	E888	01
1	44548412	05/03/1947	16/02/2003	NMO	1	21/02/2003	(null)	(null)	01
1	47398879	05/03/1947	14/03/2003	NMO	2	25/03/2003	(null)	(null)	01
2	44752387	09/07/1926	06/03/2003	COT	1	08/03/2003	(null)	(null)	01
1	44548416	05/03/1947	03/01/2003	NMO	1	09/01/2003	(null)	(null)	01
1	44548431	05/03/1947	24/01/2003	NMO	1	30/01/2003	(null)	(null)	01
1	44548443	05/03/1947	09/02/2003	URM	1	10/02/2003	(null)	(null)	01
2	44458238	11/04/1962	05/03/2003	MNC	1	07/03/2003	(null)	(null)	01
1	44864371	31/08/1960	11/02/2003	GAS	1	15/02/2003	(null)	(null)	01
2	44905500	28/04/1925	28/01/2003	MDI	1	06/02/2003	(null)	(null)	01
2	44609121	09/02/2003	09/02/2003	NNT	1	14/02/2003	(null)	(null)	01
2	44956607	19/01/1949	23/01/2003	COT	1	25/01/2003	(null)	(null)	01
2	44645546	25/08/1931	19/01/2003	MAS	1	29/01/2003	(null)	(null)	01
1	49474063	10/11/1934	26/03/2003	URM	1	26/03/2003	(null)	(null)	01
2	44875628	30/07/1972	08/01/2003	NRC	1	14/01/2003	(null)	(null)	01
1	44875632	30/07/1972	26/01/2003	NRC	1	04/02/2003	(null)	E8788	01
2	44920988	30/03/1976	26/02/2003	OFT	1	27/02/2003	(null)	(null)	01
2	49474885	04/07/1928	28/03/2003	CCR	1	29/03/2003	(null)	(null)	01
2	44744040	27/04/1950	12/02/2003	NRL	1	07/03/2003	(null)	(null)	01
2	44485993	21/03/1936	12/01/2003	RMT	1	20/01/2003	(null)	(null)	01
1	49473288	29/05/1909	24/03/2003	URM	6	25/03/2003	(null)	(null)	01
2	44965300	24/01/1923	14/01/2003	GIN	1	19/01/2003	(null)	(null)	01
2	44613660	15/02/1928	07/02/2003	MDI	1	05/03/2003	(null)	(null)	01
2	44936844	27/05/1963	22/01/2003	GIN	1	29/01/2003	(null)	(null)	01

Figure 9.3 – Results obtained

The first column (hospital_admission) shows the hospital the data comes from: 1 for Hospital Clínic of Barcelone and 2 for Hospital Virgen de las Nieves of Granada.

9.1.2.3 Discussion

These projects have proved the feasibility of using ontologies to integrate heterogeneous data adding semantic to them. In both projects UMLS was used as the conceptual reference to take the needed concepts from. Specific data can after be linked to these concepts by means of ontological relationships, mainly by subsumption. This approach is by far much more flexible and scalable than the traditional approach using relational tables to mapping codes between them. In our ontological approach adding a new classification system (for example a new CMBD) represents only adding new subclasses. In the traditional approach, adding new systems implies creating new mapping tables until a total of $\frac{n!}{2!(n-2)!}$ tables, being n the total number of classification systems.

9.2 OntoCRF

To validate the hypothesis H3 (Using an ontology-based approach it is possible to build applications addressed to health professionals), we evaluated two different aspects. On the one hand, the capability of the system for implementing applications in a real scenario. On the other one, the performance and the usability of OntoCRF from a user point of view.

9.2.1 Real implementation evaluation

OntoCRF is being marketed for several years and has been used in a variety of projects which fall in one of the following categories:

- Research projects with limited duration: a set of data, previously agreed, is collected and is finally analysed at the end of the project. This is the case of WAHA [WAHA] and Piscina [Piscina] projects.
- Clinical registries without a predetermined end date, with a long life cycle, and where it can be needed to modify the set of data to collect along the project. Running examples of these types of projects are the registry of the “Centre de reference des maladies rares du foie” [VALID] at Beaujon hospital in Paris, the Cancer Registry [RT] and the Breast Cancer Registry [ca mama] of the Hospital Clinic of Barcelona, the Registry of the "European Forum on Antiphospholipid Antibodies" for patients with the Catastrophic AntiPhospholipid Syndrome (CAPS)[CAPS] or the Registry of the "ASIA project group" for patients with the Autoimmune/inflammatory Syndrome Induced by Adjuvants (ASIA syndrome)[ASIA].
- Implementation of clinical questionnaires.
- Non clinical applications. For example the On-line Book of Clinical Residents [LRO] at Hospital Clinic of Barcelona, an application where the residents record the activities they perform during their formative period at the hospital.

When used to register patient data, the number of cases by project varies between few hundreds to three thousands, with about 60 to 600 variables per case. The number of users by project varies between less than 10 to around 500.

In all the projects, OntoCRF has been able to meet their specific requirements and, which is more interesting, to cope with the requirements of modifications during the lifecycle of the projects. The modular architecture of the metamodel has proven its feasibility to accommodate new extensions of the system. Also, the separation of data layer and presentation layer allows the progressive addition of new functionalities as needed.

The flexibility provided by the system facilitates to have prototypes from the initial moment, which is a very valuable resource in order to work close to the physicians (also customers in this case). From the very beginning of the project, key users have stuff to work with, and it is even possible to make on-line modifications and check their results immediately.

Discussion

The focus of OntoCRF is to assist to data collecting in clinical and research studies, automatizing the process as much as possible and minimizing the technical knowledge required from the final users for the creation and management of new studies. In particular, we provide an automatic system for dynamic creation of webs driven by ontologies and with additional tools for the extraction and analysis of the data.

Our system, unlike other solutions, does not work with triples or RDF graphs; it works with ontologies and more particularly with those represented in OWL. Ontologies are stored in a relational database directly in OWL, following a schema-aware approach. This hugely eases the querying process, since there is no OWL-SQL mapping needed. For instance, to retrieve the classes the system has just to access to the "class" table. Further logic is not necessary and all can be done through simple SQL queries. Since we have to deal with very large

ontologies, performance was a critical feature from the beginning and this OWL-driven approach achieved our efficiency requirements, whereas other systems failed.

OntoCRF demonstrates that an ontology-based approach is more flexible and efficient to deal with complexity and changing requirements than a traditional system, facilitating the engineering of clinical software systems. First of all, the application development phase is reduced to only analysis and design. The availability of prototypes from the very beginning, and the facility to apply changes, make OntoCRF an extremely useful tool to check the requirements and the solutions proposed. These facts imply a very important drop in costs and time with their consequent savings.

Secondly, differences between applications are reduced to their conceptual model. Therefore, the same infrastructure can be used for different projects, taking advantage of scale economy. All the projects implemented until now are sharing the same hard and soft infrastructure. The only difference between them is in the content.

At the conceptual level, some elements or models can be reused in different projects, so homogeneous criteria and conceptual models could be established inside an organization. Concepts as “patient”, “clinical manifestations”, “lab results”, etc. are very common in different projects, so its definition can be easily shared and extended as needed.

The use of ontologies provides the ability to manage data structures declaratively, thus focusing the design on the conceptual aspects and not on the technical issues. Making an ontological analysis of an application allows to focus on a higher abstraction level and to concentrate on the domain aspects, thus helping researchers to clarify the implicit knowledge to manage. This is a very

important point to guarantee de quality of the models produced. Moreover, the communication between designers and users is established at a conceptual level. This facilitates leaving aside technical discussions that often contaminate the conceptual analysis in other approaches.

Moreover, ontologies assure that data and knowledge used in the project remain very well documented. Documentation is usually left aside, or remain incomplete, in these kind of projects.

Since the solution allows the modification of the underlying schema of the data, some measures are needed to guarantee the consistency of the instances. Problems could mainly arise if trying to modify or delete classes or properties. The first security level is provided by Protégé, which does not allow performing some actions that could leave the ontology in an inconsistent state. This is the case when trying to delete a class that has instances. The rest of cases should be solved by the specification of editorial policies. When a project is running, deleting a class or property could be replaced by setting a deprecated flag on the resource. Nevertheless, in the database data are never physically deleted, only a delete flag is used, so preventing the loss of data by mistake.

Regarding the use of OWL and Protégé we consider it as a good choice. The expressivity power of the language was adequate to cover the requirements of the projects in which OntoCRF was used. Moreover, it eases the interchange and reuse of models. The use of OWL allows adding reasoning capabilities in the future, a very promising line to explore. On the other hand, this choice impose the availability of qualified professionals in these technologies, so the appropriated formative plans.

Although the system is mainly used in health related projects, the model is totally independent of the domain, so it would be suitable to gather data in

whatever context. In fact, some projects implemented with OntoCRF are not about clinical information but about management-related data. In general, if it is possible to model the data with OWL, it is possible to use OntoCRF.

9.2.2 Performance evaluation

Using the system mainly implies to navigate through the menu to access the different forms to add or modify specific data. The transaction to save data in OWL-DB is almost instantaneous, and mainly depends on traffic in the net, whereas showing the different forms depends on the data to be retrieved.

To evaluate the performance of the system perceived by the user, we measured the time required to show the forms of an application.

Method

We choose one of the more complex running projects to perform the evaluation. It is a clinical registry about patients affected by breast cancer. The experiment consisted in measuring the time required to show each one of the forms of the application, from the act of clicking the corresponding menu item until data are showed in the screen to the user. The forms used in the experiment, and the number of properties of each one, were the following:

- Patients list: it is a list with 5 properties and more than 2,000 patients in the list.
- Patient: a form with 18 properties.
- Cancer list: it is a list with 11 properties and 1 or 2 elements in the list.
- Initial cancer: a form with 33 properties.

- Tumour: a form with 13 properties.
- Cytology: a form with 19 properties.
- Breast biopsy: a form with 28 properties.
- Axillary echography: a form with 4 properties.
- Axillary cytology: a form with 20 properties.
- Axillary biopsy: a form with 21 properties.
- Tumour surgery: a form with 39 properties.
- Axillary surgery: a form with 29 properties.
- Radiotherapy: a form with 32 properties.
- Metastasis: a form with 8 properties.
- Locoregional recurrence: a form with 3 properties.
- Systemic treatment: a form with 17 properties.
- Monitoring: a form with 5 properties.

In each form there are properties of several types: single cell, radio button, combo box, check box, etc.

Analysis

Table 9.2 shows the experiment results for each form using two different web browsers, Firefox and Chrome. The time is measured in seconds with a precision of one second.

Form	Number of properties	Time using Firefox	Time using Chrome
Patients list	5	2	1
Patient	18	2	1
Cancer list	11	2	1
Initial cancer	33	2	1
Tumour	13	3	1
Citology	19	3	1
Breast biopsy	28	2	1
Axillary ecography	4	2	1
Axillary citology	20	2	1
Axillary biopsy	21	2	1
Tumour surgery	39	3	1
Axillary surgery	29	3	1
Radiotherapy	32	2	1
Metastasis	8	2	1
Locoregional	3	1	1
Systemic treatment	17	1	1
Monitoring	5	2	1

Table 9.2 – OntoCRF performance evaluation: Time in seconds required showing forms

The results shown in the table for the Chrome browser are the very same for each form, does not matter the number of properties to show. In the case of the Firefox browser there are some differences, but there is not a strong correlation between the number of properties and the time required to show the data. It seems that the browser used, and perhaps the traffic in the net, have more influence that the number of properties to show. In whatever case, the time required is quite acceptable for a web application.

9.2.3 Usability evaluation

In order to evaluate the usability of OntoCRF, the System Usability Scale (SUS) score was selected [21]. Developed in 1986 by Digital Equipment Corporation, it has been a simple method to have a first impression of the appropriateness of software developments under the point of view of the end users. It consists in a questionnaire with 10 items. The items are the following:

1. I think that I would like to use this system frequently
2. I found the system unnecessarily complex
3. I thought the system was easy to use
4. I think that I would need the support of a technical person to be able to use this system
5. I found the various functions in this system were well integrated
6. I thought there was too much inconsistency in this system
7. I would imagine that most people would learn to use this system very quickly
8. I found the system very cumbersome to use

9. I felt very confident using the system

10. I needed to learn a lot of things before I could get going with this system

The answer to each item can be a value among 1 and 5. One means “Strongly disagree” and five is “Strongly agree” with the meaning proposed by the item. The results are computed following an algorithm which gives a unique result named SUS Score with a value in the range from 0 to 100.

The data from the filled questionnaires were entered in a database and analyzed using IBM SPSS 21 Statistical Package.

A survey was distributed to a sample of 35 OntoCRF active users who used the system in a daily basis. Nineteen users (54,3 %) answered the questionnaires. Data were introduced in a database and the SUS Score computed. The results are displayed on Figure 9.4.

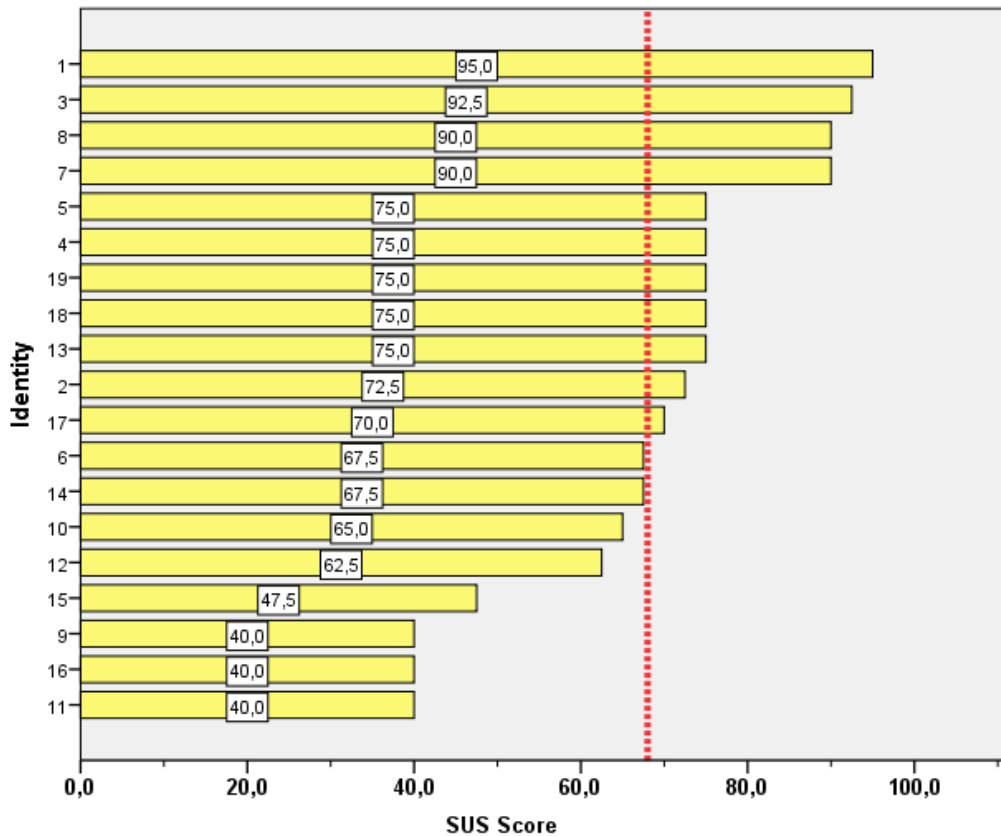


Figure 9.4 - Results of the computed SUS score, by each respondent. Results are displayed in ascending order. Dotted line marks a score of 68

Eleven of the respondents (58 %) computed a global SUS score over 68 which is recognized as “above the average” [82]. According to Bangor et al., it’s possible to grade over a curve based on the distribution of all scores in relationship with their quartile position. Four users (21 %) consider the solution to achieve an A grade (Excellent), five of them (26 %) gave a C grade (Good), six users (32 %) gave a D grade (Pass) and finally, four (21 %) rated the solution with an F grade (Fail) [9].

Related users complains about the daily working of the system, Table 9.3 shows complains received by helpdesk support during the period 2013 – 2015.

Project	Customer	Users	COMPLAINS			
			2013	2014	2015	Main complains
INTERNATIONAL ASIA SYNDROME REGISTRY	Hospital Clinic Barcelona	10	0	5	0	Slow functioning, data extraction
INTERNATIONAL CAPS REGISTRY	Fundació Clinic	257	0	0	0	
LIBRO DEL RESIDENTE ONLINE	Hospital Clinic Barcelona	519	0	14	8	Connection problems, password forgotten
MEDIOAMBIENTE Y PISCINAS	Fundació Clinic	6	1	0	0	
REGISTRO DE CANCER	Hospital Clinic Barcelona	6	0	0	3	Connection problems
BREAST CANCER REGISTRY	Fundació Clinic	24	1	9		Connection problems, slow functioning, data extraction
VALID REGISTRY	Hospital de Beaujon	13	0	0	1	Data extraction
WAHA	Fundació Clinic; Lomalinda University	43	2	8	4	Slow functioning, data extraction
TOTAL			4	36	16	

Table 9.3 – Complains received by helpdesk support

Discussion

From the usability study it can be concluded that OntoCRF is well accepted by nearly 60% of its users, who consider the solution globally above of the range. But in a more detailed look at the data, a high fragmentation is shown conforming four groups with a very different perception of usability, from the best grade of “Excellent” to the worst as “Fail”. One explanation for such discrepancy could be a different misunderstanding of the product which is under evaluation. OntoCRF has two components: a portal (developed using LifeRay®) customizable by the administrator of each different community, and a Database access for collecting the data. Moreover, OntoCRF is conceived as a full service in the cloud. Therefore, many different factors and user-experiences can be interposed in the routine operation. The SUS score was developed in 1986, when many software solutions were developed for mainframe use or in a client-server environment. At present, widespread Internet usage interposes many more layers between the user-interface and the physical data-repositories. With such scenario, we need to address much better to the users what is going to be measured with the SUS score tool and perhaps developing new tools better suited for such new systems architecture. Nevertheless, in order to improve OntoCRF it is required further usability studies including some specific questions to have better information about the reasons of a low grading by some users.

From complains received by helpdesk support during the period 2013 – 2015 it can be concluded that OntoCRF is a solid system. Practically there are not reported complains, and the very few reported are mainly related with connection problems influenced by traffic in the net.

9.3 OntoCR

As seen in Chapter 3, two hypothesis of this research were related to semantic interoperability:

H4. Standard archetypes can be used to build clinical applications.

H5. Modeling clinical information using ontologies, archetypes and controlled vocabularies is a suitable method to communicate clinical information between healthcare settings maintaining the semantic of the information.

To validate such hypothesis a complete evaluation cycle was performed using OntoCR: the creation of an archetype, the creation of a simple test application, and the communication of standardized extracts to a normalized repository [81].

We designed an archetype to gather certain basic information and the clinical stage of breast cancer samples for a tumour bank. Figure 9.5 shows a mindmap representation of the archetype.

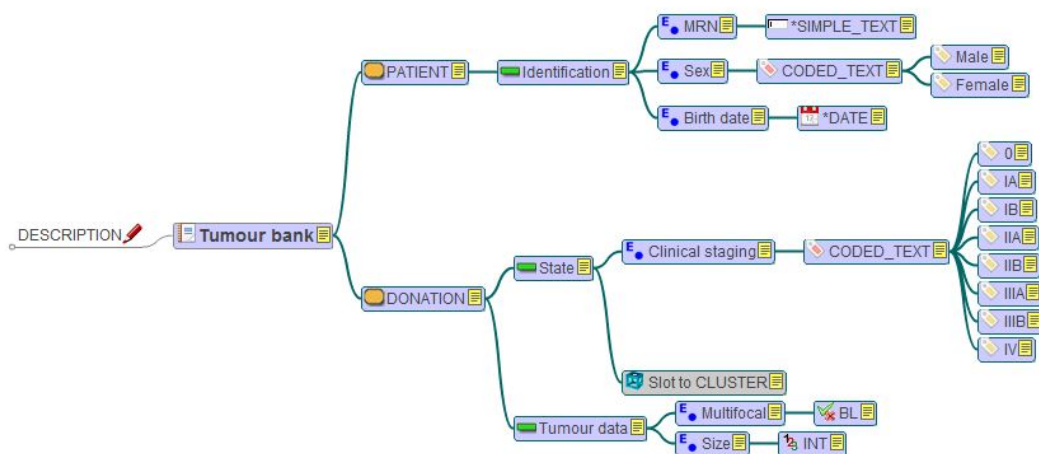


Figure 9.5 – Mindmap representing the tumour bank archetype

The archetype is a *COMPOSITION* with two *SECTIONS*:

- “Patient”. This *SECTION* has an *ENTRY* “Identification” with demographic data about the patient.
- “Donation”. This *SECTION* has two *ENTRIES*:
 - “State”. This *ENTRY* has an *ELEMENT* with the clinical staging and a *CLUSTER SLOT* which include a predefined archetype, the CEN-EN13606-CLUSTER.tnm_staging_7th-breast.v3.
 - “Tumour data”. This *ENTRY* has two more characteristic about the tumour.

The archetype was designed using LinkEHR [LinkEHR], which produces an ADL version of the archetype. Figure 9.6 partially shows the archetype in ADL format, the clinical stage of a tumour.

```

archetype
  CEN-EN13606-COMPOSITION.Tumor_Bank.v1

concept
  [at0000]
  -----
  -----

SECTION[at0004] occurrences matches {0..*} matches { -- TUMOR
  members existence matches {0..1} cardinality matches {0..*; unordered; unique} matches {
    ENTRY[at0014] occurrences matches {0..*} matches { -- Clinical stage
      items existence matches {0..1} cardinality matches {0..*; unordered; unique} matches {
        ELEMENT[at0041] occurrences matches {0..*} matches { -- Stage
          value existence matches {0..1} matches {
            CODED_TEXT[at0086] occurrences matches {0..1} matches { --
CODED_TEXT
  codedValue matches {
    CD[at0089] occurrences matches {0..1} matches { -- 0
      codeValue existence matches {0..1} matches {*}
      codingSchemeName existence matches {0..1} matches {*}
      displayName existence matches {0..1} matches {*}
      qualifiers existence matches {0..1} cardinality matches {0..*; unordered;
unique} matches {*}
      mappings existence matches {0..1} cardinality matches {0..*; unordered;
unique} matches {*}
    }
    CD[at0092] occurrences matches {0..1} matches { -- IA
      codeValue existence matches {0..1} matches {*}
      codingSchemeName existence matches {0..1} matches {*}
      displayName existence matches {0..1} matches {*}
      qualifiers existence matches {0..1} cardinality matches {0..*; unordered;
unique} matches {*}
      mappings existence matches {0..1} cardinality matches {0..*; unordered;
unique} matches {*}
    }
    CD[at0093] occurrences matches {0..1} matches { -- IB
      codeValue existence matches {0..1} matches {*}
      codingSchemeName existence matches {0..1} matches {*}
      displayName existence matches {0..1} matches {*}
      qualifiers existence matches {0..1} cardinality matches {0..*; unordered;
unique} matches {*}
      mappings existence matches {0..1} cardinality matches {0..*; unordered;
unique} matches {*}
    }
  }
  -----
  -----

```

Figure 9.6 – Representation of the clinical stage of a tumour in ADL format.

OntoCR includes a software tool to translate ADL archetype into its OWL version, in accordance with the OntoCR meta-model. Figure 9.7 shows this OWL representation in OntoCR. Ovals with thick border represent meta-classes, the other represent classes. The rectangle represents instances of the class Clinical stage. Arrows with closed head represent the relation subClassOf, arrow with opened head represent the relation instantiation. The dashed arrow represents the range of the property at0041 (Stage). The clinical stage of the tumour (node at0041 of the archetype) is modeled as an OWL property which is an instance of *ELEMENT_def*. At the same time, this property is made an instance of ComboBox, to be represented graphically. This last action is performed manually, as archetypes do not carry any information about how to be represented in a GUI.

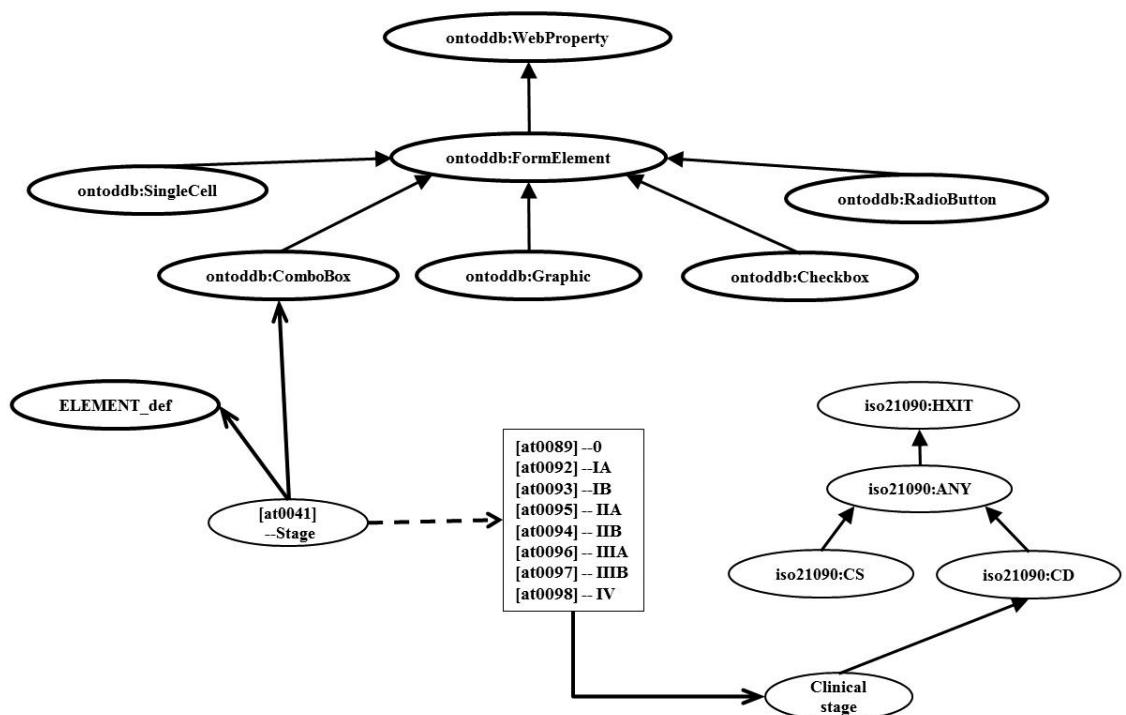


Figure 9.7 – Representation of the clinical stage of a tumour in OWL format.

Once the archetype uploaded in OntoCR it can be managed as whatever other OntoCRF application. Figure 9.7 shows how the clinical stage of the tumour (node at0041 of the archetype) is made an instance of ComboBox, to be represented graphically. This last action is performed manually, as archetypes do not carry any information about how to be represented in a GUI.

Figure 9.8 shows a web form automatically built from this representation. Once the archetype is uploaded into the system, a simple application accepts specific patient data. It is worth noting that data are directly stored as archetype instances in real time. No other conversions are needed.



Figure 9.8 – Example of application in OntoCR

In the ADL, the clinical stage is defined as an ELEMENT, node [at0041], whose values are restricted to a value set defined as several CODED_TEXT. In

OntoCR, the clinical stage is defined as a property which is, at the same time, an instance of *ontoddb:ComboBox* and *ELEMENT_def*, and which has, as its domain, the *Clinical stage* class. *ComboBox* provides properties to be represented in the GUI. *ELEMENT_def* provides properties to be considered a block of CEN/ISO 13606. The class *Clinical stage* is defined as being a subclass of the class *iso21090:CD* and has, as instances, the different allowed values.

An extraction tool has been developed to produce CEN/ISO 13606 extracts in XML format. For the purpose of validation some fictitious data were introduced and some extracts were sent to the CEN/ISO 13606 repository built by the Instituto de Salud Carlos III [62]. They were correctly validated, successfully uploaded and incorporated to the repository together with extracts coming from other organizations, produced by systems from several providers and created according to different archetypes. The idea is to have a repository of normalized and semantically interoperable information, so new research lines could be started in order to extract new knowledge by means of data mining and machine learning techniques.

Discussion

As pointed out in Chapter 4, ontologies, terminologies, information models, and data structures have overlapping functionalities. Reference models and data structures are themselves a kind of information model. Whilst this situation does not cause any problem, the consideration of some resources as knowledge models, terminologies and information models at the same time is confusing.

This situation is very clear when analysing SNOMED CT, which is mainly a clinical terminology. But SNOMED CT is not merely a list of terms; SNOMED CT defines concepts, several terms by concept, concept attributes, and

relationships between concepts, defining a specific knowledge model too. This fact makes that SNOMED CT was used as the knowledge model in some projects [56]. Although could exist a strong link between ontologies and terminologies, our belief is that they are not the same. The completeness and quality of such implicit knowledge models have been questioned [84]. In this sense, the SemanticHealth report [92] says that SNOMED CT, at this moment, can be used only as a controlled vocabulary and coding system, but it is not useful as semantic source. Similar considerations can be made about data structures and information models. Whatever resource pertaining to these categories has an implicit knowledge model, but cannot be considered as knowledge models on its own.

The system is, at this moment, at a prototype phase, but so far results are encouraging. Uploading an archetype results in an ontology with a twofold representation:

- a conceptual representation of the clinical concept subject of the archetype. The ontology, seen from this perspective, provides a semantic representation that can be used to retrieve data, and to be linked with other ontologies by describing the domain, etc.
- a representation of the archetype in terms of the reference model capable of accepting patient data as direct instances. This allows for communication of the data as CEN/ISO 13606 extracts.

Furthermore, in OntoCR an archetype can be built from scratch. OntoCR has been evaluated as a representative clinical information modeling tool [60]

Populating the archetype only requires the presentation layer, a process that is performed manually by editing the ontology. Because the system is using OntoCRF, both the storage and the user interface are obtained automatically.

The decision to follow the dual model approach of CEN/ISO 13606 has proven to be a good choice. In OntoCR, an indefinite number of archetypes can be merged (and reused) to build an application. Similar results could be achieved by using openEHR [openEHR]. Our selection of CEN/ISO 13606 is supported by the fact that it is an official and free standard, not subject to the rights of a third-party, and the fact that the Spanish Ministry of Health is in the process of creating CEN/ISO 13606 archetypes to be used at a national level [MSSSI].

The use of OntoCRF infrastructure maintains data storage totally independent of content specification. In OntoCR, we use OntoCRF infrastructure to implement the complete information architecture, not only to represent archetypes. This approach ensures greater flexibility and extensibility capabilities, which are necessary requirements of applications in the ever-changing medical field. An alternative is to translate the archetype specifications to their SQL counterparts to record instance data, a process which has several drawbacks. First, an extra step is needed to transform the archetype specification to a relational implementation. Second, this approach leads to a very complex database with a huge number of tables, and is thus difficult to maintain. Third, future modifications of the clinical model may involve modifications to the relational model, which may be difficult to implement in a system with instance data. Fourth, some characteristics of CEN/ISO 13606, such as hierarchies and nested elements, do not fit in well with the relational model and are extremely difficult (not to say impossible) to represent.

The performance of the system when populated in a real scenario remains an open question. Evaluations of OntoCRF [52] showed a linear behaviour when uploading and downloading the entire ontology. User interaction with the system in other projects using OntoCRF seems no different than with other systems.

Nevertheless, further work needs to be done to evaluate the system in a real scenario.

Although 13606 AM allows for the possibility of coexistence of different types (text or coded value) for an attribute, we think that to allow for this when designing an archetype would be a poor design choice. When implementing such an attribute in OntoCR, use of a specific type can be forced or, if it is desirable to have both available, two different attributes can be created, one for each type.

We believe the link between 13606 RM and 13606 AM to be very weak. The link is provided by the property `C_OBJECT.rm_type_name:String[1]`. When instantiated as a node of an archetype, this property should contain the name of the RM type that the node corresponds to, for example “COMPOSITION”, “ENTRY” or “CLUSTER”, but there are no real restrictions as to the value one can use. In the proposed model, there is an ontological link between both models.

ISO 21090 includes the CD.CV data type to represent coded values. This class includes a set of properties to identify the code system to which the code used belongs, but the value of these properties are strings, so the coding system itself is not conceptually represented. In OntoCR, each vocabulary is represented by a class.

The use of ontologies has demonstrated to be a very powerful solution to model the used standards. Extending OntoCRF to create OntoCR has mainly been a conceptual project, involving analysing the standard and modeling the remaining ontologies as required. In addition, working with ontologies moves the design to the conceptual level, which is more appropriate than technical discussion when modeling clinical concepts.

As mentioned before, current ISO 21090 and CEN/ISO 13606 standards focus on the communication of clinical information and no common semantic layer is assumed. For this reason, the information model is overloaded with elements that actually belong to a conceptual model. As noted in [69], the use of ontologies could solve some major clinical modeling issues, such as whether to put information in the information model or in the terminology model, and how to integrate iso-semantic models. Expressing both the information model and the terminology model in an ontology can help to avoid conceptual overlapping, and thereby facilitate its integration and lead to simpler and clearer archetype design. Moreover, this approach can facilitate management and user navigation in clinical archetype repositories [2]. The proposed system can be enriched as much as needed by integrating ontological representations of other standards, such as CEN/ISO 13940 [ISO 13940], or referencing existing ontological resources [66]. This would be a way of providing a clear ontological commitment for clinical models [57] and formally specifies how information model instances relate to clinical entities. This approach could have a direct consequence: the possibility of simplifying data type and archetype specifications. At the time of writing, OntoCR is not rooted in any upper-level ontology and, thus currently lacks a clear ontological commitment.

Ontologies have been considered promising for decades [10, 25] and now can be envisioned as a solution for common problems in the clinical domain, both as means of heterogeneous data integration [12, 24, 98] and for adding greater cognitive support to applications [34, 80].

Regarding the use of OWL, the expressivity power of the language [41] was adequate to cover the requirements of the standards, and Protégé [Protégé] has proven to be a good tool for ontology editing. The use of OWL allows, to some extent, for the automatic validation of models produced [58]. The addition of

reasoning capabilities in the future is a very promising avenue to explore. Reasoning over instance data could provide new knowledge, for example identifying repeated patterns, and the integration with domain ontologies could facilitate clinical checking as drug interactions, drug indication, etc.

The use of OWL has additional advantages. Firstly, it enables users to reuse available knowledge resources [78] and to link them with archetype definitions, thereby adding knowledge to the system. Secondly, there is increasing use of OWL as formalism for representing clinical models [44, 46, 58, 76, 94]. Moreover, OWL is a standard with wide support among the Semantic Web community [OWL activity]. Consequently, tools developed by the Semantic Web community can be directly applied to these models.

Despite all these efforts, there is really no practical experience, as far as we know, of using currently these kinds of models. The most similar work is the proposal of Tao et al [94] to represent the clinical element model (CEM) specification, an information model designed for representing clinical information in EHR systems. They used a three-layer model in OWL: (1) a meta-level ontology that defines the meta-representation of the CEM; (2) OWL ontologies for representing each individual CEM; and (3) patient data represented as instances of the ontologies on layer 2. The system does not include the possibility to define the graphical user interface. Legaz-García et al [43] propose an archetype management system that uses OWL ontologies to represent CEN/ISO 13606 archetypes. This approach enables the semantic management of archetypes providing interesting functionalities as the transformation of archetypes between standards or their validation. Their proposal does not include how to use the archetypes to record patient data. There are some proposals that use a relational representation. Austin et al [7] developed an EHR server inside a relational database using CEN/ISO 13606 as the

information basis for the design of the server. In this case, the choice of a relational representation imposes limits that impede the representation of many features of the standard. Wang et al [99] propose generating relational databases mapping openEHR archetypes to data tables, which implies an extra translation layer. We believe that using a direct relational representation is a less flexible approach and might be difficult to manage versioning.

9.4 Summary

In this chapter we described the evaluations performed for validating the hypotheses of this thesis. For that purpose, we run multiple experiments on each technical component of OntoEHR: OntoDDB, OntoCRF and OntoCR, and analysed the results. These results promisingly confirm the adequacy of the technical infrastructure of OntoEHR in their current state and validate our hypotheses. Nevertheless, it lacks the implementation of a real medical record according to the proposed model, something we expect to see in the future as described in the following chapter.

Chapter 10

Conclusions and Future Work

In this thesis we present OntoEHR, a conceptual architecture for a new semantically interoperable and knowledge enriched EHR system, focused on the clinical process and driven by ontologies. The main goal is to provide value-added assistance to healthcare professionals in the decisions who must exercise in their daily practice, for both primary and secondary use of clinical data. With the aim of building information systems oriented towards the healthcare professional the first objective was:

O1. To assist healthcare professionals gathering and recording structured clinical information and its reuse, as electronic medical record, epidemiological registries and clinical research.

According to the analysis presented in Chapter 1, the better way of providing support for healthcare professionals is building information systems that truly implement the clinical process. The clinical process transform a health state as input into another health state as output. Health states being abstract entities are perceived as patient health problems, which constitute the lynchpin on which to organize the patient care in the care model proposed.

On the other hand, clinical data should be recorded in a standardized and structured way, to facilitate its reuse.

As has been indicated in chapter 2, healthcare tends to be provided through different healthcare entities, making it harder the conceptual consolidation of patient's clinical situation. In this scenario with multiple healthcare organizations involved, semantic interoperability is essential for continuity of care. For that reason, our second objective was:

O2. To facilitate the continuity of care between health facilities.

To achieve this goal, the care model proposed is conforming with an international standard defining the system of concepts to support continuity of care, the CEN/ISO 13940 standard. Moreover, we propose another standard, CEN/ISO 13606, to communicate extracts of clinical information among different clinical information systems.

Current information systems cannot accommodate change easily, but change is a constant in the medical domain, so our third objective was:

O3. To design a system that is at the same time flexible, solid, and efficient, capable to deal with the complexity and change of clinical domain.

A main characteristic of OntoEHR is that the specification of an implemented system is fully and explicitly declared in a set of ontologies. So, changing the system becomes an ontology edition problem, not involving database design or programming. At the same time, the use of a relational storage provides the needed solid base and efficiency.

Medicine is a science based on knowledge but the increasingly available knowledge is not integrated in current systems. Our fourth objective was:

O4. To seamless incorporate clinical knowledge into the clinical information systems.

The use of ontologies as the core that drive all the system, allows to seamless incorporate clinical knowledge from external sources as new components. Establishing relationships between added knowledge and already existing concepts into the system is a way the system to learn.

We discuss in the following sections the main contributions of this research and review if our initial hypotheses were verified by the performed experiments. Then, we provide an outlook for some future lines of work. Finally, the main conclusions are presented.

10.1 Main Contributions

In this thesis we propose a novel approach to build EHR systems which leads to our main contribution, OntoEHR, a conceptual architecture for a new semantically interoperable EHR (**C1**). OntoEHR is conceptually based on the representation of the clinical process and aims to assist healthcare professionals both for primary and secondary use of clinical information. OntoEHR is technically based on the declarative specification of the system by ontologies, so being independent the conceptual and technical layers. Both conceptual and technical solutions, are based on the extensively use of international standards in health domain.

The first component we created for OntoEHR was OntoDDB, a framework for the definition and storage of ontologies (**C2**). Being capable to store OWL ontologies, OntoDDB can be used to build both knowledge servers and data repositories. Storing the data in a relational database provides the known

advantages of a solid relational model. OntoDDB is a tool that allows the full cycle of ontology editing in the medical domain, suitable for medical experts without special training on software. This can facilitate the incorporation of medical experts to the task of system specification.

The next step was the creation of a framework to build complete applications. Therefore, we developed OntoCRF, a framework for the definition, modeling, and instantiation of data repositories (C3). For data storage and ontology edition OntoCRF uses OntoDDB, and defines a metamodel for the specification of user applications. OntoCRF is a complete framework to build data repositories since includes design of the system, storage and GUI. The combination of ontologies and relational technology provides a system that is at the same time flexible and solid. The ontology-based approach is more flexible and efficient to deal with complexity and change than traditional systems.

OntoCRF does not require very skilled technical people to make a new project, easing the engineering of clinical software systems. Moreover, the reduction of the development phase implies a very important drop in costs and time with the consequent saving. Furthermore, as the same infrastructure can be used for different projects, there is no need to dedicate specific equipment for each new project.

At the conceptual level, the ontological analysis of applications allows to concentrate on the domain aspects, helping researchers to clarify the implicit knowledge to manage, and to facilitating the communication between designers and users. As some concepts are very common in different projects, the models can be reused. On the other hand, ontologies assure that data and knowledge used in the project remain very well documented.

Once we had a framework to build user applications, the next step was to achieve semantic interoperability with other clinical information systems. For this reason we developed OntoCR, a semantically interoperable clinical repository, based on ontologies, and conforming to CEN/ISO 13606 standard (C4). OntoCR demonstrates that is possible to build a native CEN/ISO 13606 repository for the storage of clinical data. Our approach has been to extend an existing framework for the development of clinical data repositories driven by ontologies: OntoCRF. The similar approach of OntoCRF and CEN/ISO 13606, a dual model separating information and knowledge, establishes a natural way to do it, by adding the conceptual models of CEN/ISO 13606 to the OntoCRF metamodel. Moreover, the proposed system can be enriched as needed by integrating ontological representations of other standards or referencing existing ontological resources.

Furthermore, we have demonstrated semantic interoperability of clinical information using CEN/ISO 13606 between a sender and a receiver, which is the result of independent developments. This is a pioneering experience at an international level.

In order to really implement an EHR system it is not enough with the mentioned components. We needed a model of EHR which was compliant with this framework. We have defined a Problem Oriented Medical Record model, focused on the representation of the clinical process (C5). The model proposed is conforming with the standard ISO 13940, facilitating in this way the continuity of healthcare and organizational interoperability between centers.

Finally, we successively created an implementation of each of the components. As result, we obtained a tool that implements the above models, thus allowing the recording of clinical data in a new EHR system (C6). Furthermore, OntoCRF is being marketed and used in several national and international projects.

10.2 Hypotheses verification

During this research we performed several experiments (see Chapter 9) to evaluate our contributions and determine if our initial hypotheses were successfully verified.

First, we evaluated the possibility of using ontologies as operative databases in the clinical domain on two different perspectives: the adequacy of the system and the capability of semantically integrating heterogeneous data.

To evaluate the adequacy of the system we used 7 different ontologies to test if the time needed to store and retrieve an entire ontology was appropriate. The results obtained showed that its behaviour was quite linear. In general, the upload and download times are proportional to the number of statements. The greater complexity of some ontologies, expressed by a higher proportion of classes and properties in relation to the number of instances, involves a slight penalty.

The capability of integrating heterogeneous data was tested in two different projects. In SCOPE project, a set of scientific articles were represented in an ontology integrated with concepts extracted from UMLS. In INBIOMED project, an ontology was used to integrate the CMBD coming from two different hospitals. Both projects proved the feasibility of using ontologies to integrate heterogeneous data adding semantic to them. Furthermore, this approach is by far much more flexible and scalable than the traditional approach using relational tables to mapping codes between them.

In conclusions, these experiments verify that OntoDDB can be used as storage solution in OntoEHR, thus validating hypotheses **H1** and **H2**.

We evaluated three different aspects of OntoCRF to build applications: the capability of the system for implementing applications in a real scenario, the

performance and the usability of OntoCRF from a user point of view. Related the capability of the system to build real applications, OntoCRF has been implemented in a variety of projects and in all of them has been able to meet their specific requirements and, which is more interesting, to cope with the requirements of modifications during the lifecycle of the projects.

To test the performance of OntoCRF perceived by the user, we measured the time required to show the forms of an application. The results showed an acceptable performance, more influenced by the browser used and the traffic in the net.

Finally, in order to evaluate the usability of OntoCRF, a survey was distributed among active users, using the System Usability Scale (SUS) score. The results showed that OntoCRF is well accepted by nearly 60% of its users, who consider the solution globally above of the range.

The results showed that using an ontology-based approach it is possible to build applications addressed to health professionals, so validating hypothesis **H3**.

Lastly, we needed to evaluate the semantic interoperability capability of the system. To do that, a complete evaluation cycle was successfully performed using OntoCR: the creation of an archetype, the creation of a simple test application, and the communication of standardized extracts to a normalized repository. With this experiment we validated that standard archetypes can be used to build clinical applications (**H4**) and that Modeling clinical information using ontologies, archetypes and controlled vocabularies is a suitable method to communicate clinical information between healthcare settings maintaining the semantic of the information (**H5**).

10.3 Future Work

In this section we identify different lines for future work to overcome the current limitations of the system and to extend its capabilities.

- *Adding query functionalities to the final user.* In this thesis we focused on data and process representation. Although it is possible to perform some data extraction and ontology querying, more user friendly tools can be developed. We have already worked on the integration of existing data query tools with OntoCRF in order to provide query functionalities to the final user. Nevertheless, including SPARQL capabilities would greatly improve the exploitation of stored data.
- *Deeper use of existing ontologies and tools.* At this moment ontologies are being used in OntoCR as a matter of data modeling tool, so the use of already existing ontologies is a natural step. Moreover, to apply automatic reasoning to data gathered in a project, which are integrated with external ontologies, could provide interesting benefits.
- *Providing an ontological commitment to OntoCR using an upper-level ontology.* The use of BioTopLite2 [BioTop] for this purpose may be considered in the future as a potential solution, representing OntoCR classes under “Information object”. This would help to separate the representation of information and what it refers to [85], something usually mixed and confused in current EHR systems.
- *Testing the system in a real scenario.* Although the different technological components of OntoEHR have been successfully tested, the full model is at this moment a proof of concept focused on design aspects. The model of EHR proposed has not been evaluated yet. It would be very interesting to perform some pilot implementing the POMR proposed.

- *Enhance the uploading of archetypes and generation of extracts.* We are currently using software tools to allow the uploading of archetypes to the system and the generation of extracts. This software tools can be easily transformed to web services.
- *Transforming data from legacy systems.* Current clinical information systems contain a huge amount of structured and unstructured data difficult to query. Transforming these data to OntoEHR would improve their usability.
- *Extracting new knowledge.* The possibility of using the repository of normalized information as a source for new knowledge also worth to be studied.

10.4 Conclusions

The main conclusions of this thesis are the following:

- Ontologies are feasible as operative databases in the clinical domain.
- Using an ontology-based approach it is possible to build at the same time flexible, solid, and efficient applications addressed to health professionals.
- The integrated use of ontologies, archetypes and controlled vocabularies is a suitable method to provide semantic interoperability when communicating clinical information between different healthcare systems.
- The use of ontologies allows incorporating clinical knowledge into the clinical information systems.
- With the results achieved in this thesis we set the foundations to develop a Problem Oriented Model Record representing the clinical process,

conforming with the system of concepts defined by CEN/ISO 13940 standard.

The main goal of this thesis, to provide true support to healthcare professionals in their daily practice, for both primary and secondary use of clinical data, is a core requirement to support new generations of EHR systems. This thesis provides a step forward in that direction.

Bibliography

- [1] Achour S L et al. A UMLS-based knowledge acquisition tool for rule-based clinical decision support system development. *J.Am.Med.Inform.Assoc.* 8.4 (2001): 351-60
- [2] Allones JL, Taboada M, Martinez D, Lozano R, Sobrido MJ. SNOMED CT module-driven clinical archetype management. *J Biomed Inform.* 2013 Jun;46(3):388-400. doi: 10.1016/j.jbi.2013.01.003
- [3] Anderson NR, Lee ES, Brockenbrough JS, et al. Issues in biomedical research data management and analysis: needs and barriers. *J Am Med Inform Assoc;* 2007; 14:478–88
- [4] Anhoj J. Generic design of web-based clinical databases. *Journal of Medical Internet Research;* Oct 2003; 5(4)
- [5] Asenjo MA, Bertrán MJ, Guinovart C, Llach M, Prat A, Trilla A Analysis of Spanish hospital's reputation: relationship with their scientific production in different subspecialities. *Med Clin (Barc);* 2006 May 27; 126(20):768-70
- [6] Atalag K et al. Putting Health Record Interoperability Standards to Work. *eJHI* 5(1) 2010
- [7] Austin T, Sun S, Lim YS, Nguyen D, Lea N, Tapuria A, Kalra D. An Electronic Healthcare Record Server Implemented in PostgreSQL. [J Healthc Eng.](#) 2015;6(3):325-44.

- [8] Ball MJ, Silva JS, Bierstock S, Douglas JV, Norcio AF, Chakraborty J, et al. Failure to Provide Clinicians Useful IT Systems: Opportunities to Leapfrog Current technologies. *Methods Inf Med* 47:4-7. 2008
- [9] Bangor A, Kortum Ph, Miller J: Determining what individual SUS Scores mean: adding and Adjective Rating Scale. *Journal of Usability Studies*, 2009, 4(3):114-123
- [10] Berners-Lee T, Hendler J, Lassila O. The Semantic Web. *Scientific Am.*, May 2001, pp. 34-43
- [11] Berwick D, Nolan T, et al. The triple aim: care, health, and cost. *Health Affairs*, 27: 759-769. 2008
- [12] Bettencourt-Silva J, de la Iglesia B, Rayward-Smith V. On Creating a Patient-centric Database from Multiple Hospital Information Systems. *Meth Inform Med* 2012 51:210-220. doi: 10.3414/ME10-01-0069
- [13] Bisbal J, Berry D. An Analysis Framework for Electronic Health Record Systems. *Methods Inf Med* 50(2) 180-9. 2011. doi: 10.3414/ME09-01-0002.
- [14] Bishop B, Kiryakov A, Ognyanoff D, Peikov I, Tashev Z, and Velkov R. OWLIM: A family of scalable semantic repositories. In *Journal Semantic Web*; 2011; 2(1):32-42
- [15] Paul E. Black, "data structure", in *Dictionary of Algorithms and Data Structures* [online], Vreda Pieterse and Paul E. Black, eds. 15 December 2004. Available from: <http://www.nist.gov/dads/HTML/datastructur.html>
- [16] Bodenreider O, Burgun A. Biomedical ontologies. In: Chen H, Fuller S, Hersh WR, Friedman C, editors. *Medical informatics: Advances in knowledge management and data mining in biomedicine*. New York: Springer-Verlag; 2005: 211-236

- [17] Booch G, Brown A, Iyengar S, Rumbaugh J, and Selic B. An MDA manifesto. In MDA Journal: Model driven architecture straight from the masters, chapter 11, May 2004
- [18] Bouamrane M.M., Tao C., Sarkar I.N. Managing Interoperability and Complexity in Health Systems. *Methods Inf Med* 54(1), 2015
- [19] Boulos MN, Roudsari AV, and Carson ER. Towards a semantic medical Web: HealthCyberMap's tool for building an RDF metadata base of health information resources based on the Qualified Dublin Core Metadata Set. *Med.Sci.Monit.* 8.7 (2002): MT124-MT136
- [20] Brickley D, Guha RV. Resource Description Framework (RDF) Schema Specification 1.0. W3C Candidate Recommendation. Brickley D, Guha RV, editors. 2003. <http://www.w3.org/TR/rdf-schema/>
- [21] Brooke J: SUS – A quick and dirty usability scale. In Jordan PW, Thomas B, Weerdmeester BA and McClelland IL, Eds. ; *Usability Evaluation in Industry*, pp. 189-194. London: Taylor & Francis. 1996
- [22] Butt AS, Haller A, Liu S, Xie L. ActiveRaUL: Automatically Generated Web Interfaces for Creating RDF Data. *Semantic Web 2013*
- [23] Celko J. Nested set model of trees in SQL. In *SQL for Smarties: Advanced SQL Programming*. 2^a ed. San Francisco: Morgan Kaufmann Publishers. 1999
- [24] Ceusters W, Smith B, De-Moor GJ. Ontology-Based Integration of Medical Coding Systems and Electronic Patient Records. *IFOMIS Reports* 2004. <http://ontology.buffalo.edu/medo/CodingAndEHCR.pdf>. (Accessed January 20, 2016)
- [25] Chandrasekaran B, Josephson JR. What are ontologies and why do we need them?. *IEEE Intelligent Systems*, 1999 1:20-26

- [26] Charles D, Gabriel M, Searcy T. (April 2015) Adoption of Electronic Health Record Systems among U.S. Non-Federal Acute Care Hospitals: 2008-2014. *ONC Data Brief, no.23*. Office of the National Coordinator for Health Information Technology: Washington DC
- [27] Cimino JJ, Ayres EJ. The Clinical Research Data Repository of the US National Institutes of Health. *Studies in health technology and informatics* 2010;160(Pt 2):1299-1303.
- [28] Coiera E. Building a National Health IT System from the Middle Out. . *J Am Med Inform Assoc* 2009; 16:271-273. DOI 10.1197/jamia.M3183
- [29] Das S, Srinivasan J. Database technologies for RDF. *Lecture notes in computer science*. Volume 5689: 205-221. Springer-Verlag. 2009
- [30] Gray BH, Bowden T, Johansen I, Koch S. Electronic Health Records: An International Perspective on “Meaningful Use”. *The Commonwealth Fund Publication* 28:1565, November 2011.
<http://www.commonwealthfund.org/publications/issue-briefs/2011/nov/electronic-health-records-international-use>
- [31] Grimson J. Delivering the electronic healthcare record for the 21st century. *Int J Med Inform* 2001; 64:111–127.
- [32] Gruber TR. A translation approach to portable ontology specifications. *Knowledge Acquisition*; 1993 5(2): 199-220. Academic Press
- [33] Guarino N, Oberle D, Staab S. What Is an Ontology? In *Handbook on Ontologies*, pp 1-17. Springer Berlin Heidelberg, Berlin, Heidelberg. 2009. ISBN=978-3-540-92673-3, doi: 10.1007/978-3-540-92673-3_0

- [34] Haug PJ, Ferraro JP, Holmen J, Wu K, Mynam K, Ebert M, Dean N, Jones J. An ontology-driven, diagnostic modeling system. *J Am Med Inform Assoc*. 2013 Jun;20(e1):e102-10. doi: 10.1136/amiajnl-2012-001376
- [35] Hersh WR, Weiner MG, Embi PJ, et al. Caveats for the Use of Operational Electronic Health Record Data in Comparative Effectiveness Research. *Medical care*. 2013;51(8 0 3):S30-S37. doi:10.1097/MLR.0b013e31829b1dbd.
- [36] Hovenga E.J.; Garde S. Electronic Health Records, Semantic Interoperability and Politics. *eJHI* 5(1) 2010
- [37] Hruby GW, McKiernan J, Bakken S, Weng Ch. A centralized research data repository enhances retrospective outcomes research capacity: a case report. *J Am Med Inform Assoc*; published 15 January 2013
- [38] Hsiao C-J, Hing E. Use and characteristics of electronic health record systems among office-based physician practices: United States, 2001–2013. NCHS data brief, no 143. Hyattsville, MD: National Center for Health Statistics. 2014
- [39] Especial Red Temática en Informática Biomédica: INBIOMED. *Informática y Salud* n° 46. Sociedad Española de Informática de la Salud. Madrid. Abril 2004. <http://www.conganat.org/seis/is/is46/>
- [40] Knublauch H. Ontology-driven software development in the context of the semantic web: an example scenario with Protégé/OWL. In Frankel, D.S., Kendall, E.F., and McGuinness, D.L. eds. 1st International Workshop on the Model-Driven Semantic Web; 2004 (MDSW2004)
- [41] Kola J, Wheeldin B, Rector A. Lessons in building OWL Ontology driven applications: OCHWIZ – an Occupational Health Application. *Proceedings All Hands 2007*

- [42] Lassila O, Swick RR. Resource Description Framework (RDF) Model and Syntax Specification. W3C Recommendation. Lassila O, Swick RR, editors. 1999. <http://www.w3.org/TR/REC-rdf-syntax/>
- [43] Legaz-García MC, Martínez-Costa C, Menárquez-Tortosa M, Fernández-Breis JT. Exploitation of ontologies for the management of clinical archetypes in ArchMS. Proceedings of the 3rd International Conference on Biomedical Ontology (ICBO 2012), KR-MED Series. Graz, Austria, July 21-25, 2012
- [44] Legaz-García M del C, et al. Transformation of standardized clinical models based on OWL technologies: from CEM to OpenEHR archetypes. *J Am Med Inform Assoc* 2015;0:1–9. doi:10.1093/jamia/ocu027
- [45] Lehmann CU, Altuwaijri MM, Li YC, Ball MJ, Haux R. Translational Research in Medical Informatics or from Theory to Practice. *Methods Inf Med*;47:1-3. 2008.
- [46] Lezcano L, Sicilia MA, Rodriguez-Solano C. Integrating reasoning and clinical archetypes using OWL ontologies and SWRL rules. *J Biomed Inform* 2011;44:343–53.
- [47] Li Y-F, Kennedy G, Ngoran F, Wu P, Hunter: An Ontology-centric Architecture for Extensible Scientific Data Management; Future Generation Computer Systems, Elsevier, July 2011
- [48] Lozano-Rubí R, Prat S, Echeverría T, Pastor X. Trauma registry. Research and clinical records Proceedings Book .Third European Conference on Electronic Health Care Records; 1999: 199-208
- [49] Lozano-Rubí R. Registros clínicos electrónicos y su explotación para investigación *Todo Hospital*; 1999; 162: 817-822

- [50] Lozano-Rubí R, Geva F, Saiz S, Pastor X: Relational Support for Protégé. Sixth International Protégé Workshop; 2003. Manchester.
- [51] Lozano-Rubí R, Pastor X, Lozano E. OntoDDB - Ontology Driven Database. Proc First Symp Healthcare Systems Interop; 2009: 31-38, ISSN 2174-7415
- [52] Lozano-Rubí R, Pastor X, Lozano E. OWLing Clinical Data Repositories with the Ontology Web Language. JMIR Med Inform 2014;2(2):e14
URL: <http://medinform.jmir.org/2014/2/e14/>
doi:10.2196/medinform.3023
- [53] Lozano-Rubí R, Muñoz Carrero A, Serrano Bazalote P, Pastor X. OntoCR: A CEN/ISO-13606 clinical repository based on ontologies. Journal of Biomedical Informatics 2016; 60:224-233
doi: 10.1016/j.jbi.2016.02.007
- [54] MacKenzie, SL; Yatt, MC; Schuff, R; Tenenbaum, JD; Anderson, N. Practices and perspectives on building integrated data repositories: results from a 2010 CTSA survey. J Am Med Inform Assoc 2012; 19. doi: 10.1136/amiajnl-2011-000508
- [55] Magkanaraki A, Karvounarakis G, Tuan Anh T, Christofides V, Plexousakis D. Ontology Storage and Querying. Technical Report N° 308. Greece, Foundation for Research and Technology Hellas. Institute of Computer Science. Information Systems Laboratory. 2002
- [56] Markwell D, Sato L, Cheetham E. Representing clinical information using SNOMED Clinical Terms with different structural information models, Proceedings of the 3rd international conference on Knowledge Representation in Medicine (KR-MED 2008). 2008

- [57] Martínez-Costa C, et al. Semantic enrichment of clinical models towards semantic interoperability. The heart failure summary use case. *J Am Med Inform Assoc* 2015;0:1–12. doi:10.1093/jamia/ocu013
- [58] Menárguez-Tortosa M, Fernández-Breis JT. OWL-based reasoning methods for validating archetypes. *J Biomed Inform.* 2013;46(2):304–317.
- [59] Mitchell JA, Gerdin U, Lindberg DA, Lovis C, Martin-Sanchez F, Miller RA, Shortliffe EH, Leong TY. 50 Years of Informatics Research on Decision Support: What's Next. *Methods Inf Med* 50, pp 525-535. 2011
- [60] Moreno-Conde A, Austin T, Moreno-Conde J, Parra-Calderón C, Kalra D. Evaluation of clinical information modeling tools. *J Am Med Inform Assoc* 2016;0:1-9. doi: 0.1093/jamia/ocw018
- [61] Motik B, Maedche A, Volz R. A Conceptual Modeling Approach for Semantics-Driven Enterprise Applications. *Lecture Notes in Computer Science* vol. 2519, 1082-1099. 2002
- [62] Muñoz A, Somolinos R, Pascual M, Fragua JA, Gonzalez MA et al. Proof-of-concept Design and Development of an EN13606-based Electronic Health Care Record Service. *J Am Med Inform Assoc.* 2006; 14(1):118-112.
- [63] Musen MA, Schreiber ATh. Architectures for intelligent systems based on reusable components. In *Artificial Intelligence in medicine*; 1995; 7: 189-199
- [64] Musen MA. Scalable software architectures for decision support. *Methods of Information in Medicine*; 1999 38:229-238
- [65] Musen MA. Ontology-oriented design and programming. In: Cuenca J, Demazeau Y, Garcia A, and Treur J eds. *Knowledge engineering and agent technology*; 2004. Amsterdam: IOS Press

- [66] Musen MA, Noy NF, Shah N, et al. The National Center for Biomedical Ontology. *J Am Med Inform Assoc* 2012 19:190-195. doi: 10.1136/amiajnl-2011-000523
- [67] Neuhaus F et al. Towards ontology evaluation across the life cycle. *Applied Ontology*, 2013, 8(3)
- [68] Noy NF, McGuinness D. *Ontology Development 101: A Guide to Creating Your First Ontology*.
http://liris.cnrs.fr/amille/enseignements/Ecole_Centrale/What%20is%20an%20ontology%20and%20why%20we%20need%20it.htm
- [69] Oniki TA, Coyle JF, Parker CG, Huff SM. Lessons learned in detailed clinical modeling at Intermountain Healthcare. *J Am Med Inform Assoc*. 2014; 21(6), doi:10.1136/amiajnl-2014-002875
- [70] OWL Web Ontology Language Reference. <http://www.w3.org/TR/owl-features/>
Archived by WebCite® at <http://www.webcitation.org/6KDYDWVMT>
- [71] Pan Z, Zhang X, and Heflin J. DLDB2: A Scalable Multi-perspective Semantic Web Repository. In *Web Intelligence*, IEEE; 2008; 489-495
- [72] Pastor X; Lozano-Rubí R. Representación de la información clínica e investigación sobre salud. En *Informática Biomédica*; 2004: 115; INBIOMED; ISBN 84-609-1770-3
- [73] Pathak J, Bailey KR, Beebe CE, et al. Normalization and standardization of electronic health records for high-throughput phenotyping: the SHARPN consortium. *J Am Med Inform Assoc* 2013;20: e341–e348.

[74] Payne TH, Corley S, Cullen TA, et al. Report of the AMIA EHR 2020 Task Force on the Status and Future Direction of EHRs. DOI: <http://dx.doi.org/10.1093/jamia/ocv066>

[75] Ranking de resultados de los INSTITUTOS DE INVESTIGACIÓN SANITARIA. Available at: <http://www.isciii.es/ISCIII/es/contenidos/fd-el-instituto/fd-comunicacion/fd-noticias/21Dic2012-Institutos-de-Investigacion-Sanitaria.shtml> (in Spanish)
Archived by WebCite® at <http://www.webcitation.org/6KDY23F8v>

[76] Rea S, Pathak J, Savova G, Oniki TA, Westberg L, Beebe CE, Tao C, Parker CG, Haug PJ, Huff SM, Chute CG. Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: The SHARPN project. *Journal of Biomedical Informatics* 45 (2012) 763–771

[77] Rector, A.L.; Qamar, R.; Marley, T. Binding ontologies and coding systems to electronic health records and messages. *Applied Ontology*; 2009 4(1). IOS Press

[78] Rector A, Brandt S, Drummond N, Horridge M, Pulestin C, Stevens R. Engineering use cases for modular development of ontologies in OWL. *Applied Ontology* 2012 7(2) 113-132

[79] Safran C, Bloomrosen M, Hammond WE, et al., Toward a National Framework for the Secondary Use of Health Data: An American Medical Informatics Association White Paper. *Journal of the American Medical Informatics Association* : JAMIA 2007;14(1):1-9. doi:10.1197/jamia.M2273.

[80] Saleem JJ, Flanagan ME, Wilck N, Demetriades J, Doebbeling B. The next-generation electronic health record: perspectives of key leaders from the US Department of Veterans Affairs. *J Am Med Inform Assoc* published online April 18, 2013. doi: 10.1136/amiajnl-2013-001748

- [81] Sánchez de Madariaga R, Cáceres Tello J, Muñoz Carrero A, Moreno Gil O, Velázquez Aza I, Somolinos Cristóbal R. Public electronic Health Record Platform Compliant with the ISO EN13606 Standard as Support to Research Groups. XIII Mediterranean Conference on Medical and biological Engineering and Computing 2013. MEDICON 2013. IFMBE Proceedings. ISBN: 978-3-319-00845-5 (Print) 978-3-319-00846-2 (Online) Sevilla. España 25-28 Sep. 2013.
- [82] Sauro J. Measuring Usability with the System Usability Scale (SUS). February, 2, 2011.
Accessible at <http://www.measuringusability.com/sus.php>
- [83] Schulz S, Schober D, Daniel C, Jaulent MC. Bridging the semantics gap between terminologies, ontologies, and information models. Medinfo 2010. Brisbane.
- [84] Schulz S, Cornet R, Spackman K. Consolidating SNOMED CT's ontological commitment. *Applied Ontology* 6(1), pp 1-11. 2011
- [85] Schulz S, Martínez-Costa C, Karlsson D, Cornet R, Brochhausen M, Rector A. An Ontological Analysis of Reference in Health Record Statements. Conference on Formal Ontology in Information Systems (FOIS 2014), Rio de Janeiro, Brasil, September 22-26, 2014.
- [86] Senger C, Seidling HM, Quinzler R, Leser U, Haefeli WE. Design and Evaluation of an Ontology-based Drug Application Database. *Methods Inf Med*; 2011; 50(3): 273-284
- [87] Shepherd M. Challenges in Health Informatics. *Proc 40 HICSS*;1-10. 2007.
- [88] Smith B, Brochhausen M. Putting Biomedical Ontologies to Work. *Meth Inform Med* 2010; 49(2):135-140

- [89] SNOMED-CT Starter Guide 2014. http://snomed.org/sg_gb.pdf.html
<http://www.ihtsdo.org/snomed-ct/>
- [90] Soguero-Ruiz C, Lechuga LA, Mora-Jimenez I, Ramos-López J, Barquero O, García-Alberola A, Rojo JL. Ontology for Heart Rate Turbulence Domain Applying the Conceptual Model of SNOMED-CT. *Computing in Cardiology* 2012; 39:89-92
- [91] Stephen B, Johnson Ph D, Paul T, Khenina A. Generic Database Design for Patient Management Information. *Proceedings of the AMIA Annual Fall Symposium*, 22-27. 1997.
- [92] Stroetmann V N, Kalra D, Lewalle P, Rector A, Rodrigues J M, Stroetmann K A, Surjan G, Ustun B, Virtanen M, and Zanstra PE. *Semantic Interoperability for Better Health and Safer Healthcare. Semantic Health Report*, 2009. . European Comission - Information Society and Media.
- [93] Tang P. Key Capabilities of an Electronic Health Record System. Letter Report. Institute of Medicine Committee on Data Standards for Patient Safety. Board on Health Care Services. Washington D.C.: National Academies Press. July 31, 2003
- [94] Tao C. et al. A semantic-web oriented representation of the clinical element model for secondary use of electronic health records data. *Journal of the American Medical Informatics Association* 20(3) 554-562. 12/2012
- [95] Theoharis Y, Christophides V, Karvounarakis G. Benchmarking Database Representations of RDF/S Stores. *Lecture Notes in Computer Science*; 2005; Volume 3729: 685-701. Springer-Verlag
- [96] Tudorache T, Nyulas C, Noy NF, Musen M: *WebProtégé: A Distributed Ontology Editor and Knowledge Acquisition Tool for the Web. Semantic Web journal*, 2011

- [97] Uschold M, Gruninger M. *Ontologies: Principles, Methods and Applications*. Knowledge Engineering Review 11(2). 1996
- [98] Wache H, Vögele T, Visser U, Stuckenschmidt H, Schuster G, Neumann H, and Hübner S. "Ontology-based Integration of Information - A Survey of Existing Approaches," In: Proceedings of IJCAI-01 Workshop: Ontologies and Information Sharing, Seattle, WA, 2001, Vol. pp. 108-117.
- [99] Wang L, Min L, Wang R, Lu X, Duan H. Archetype relational mapping – a practical openEHR persistence solution. BMC Medical Informatics and Decision Making. 2015 Vol 15. DOI 10.1186/s12911-015-0212-0
- [100] Weed LL. Medical records that guide and teach. New Engl J Med Volume 278. p. 593-600.
- [101] Weed LL. Representation of Medical Knowledge and PROMIS. In: Blum BI, editor. Information systems for patient care. New York: Springer; 1984. p. 83-108.
- [102] Wright A, Sittig DF, McGowan J, Ash JS, Weed LL. Bringing science to medicine: an interview with Larry Weed, inventor of the problem-oriented medical record. J Am Med Inform Assoc volume 21(6). p. 964-968.
- [103] Zhou J, Ma L, Liu Q, Zhang L, Yu Y, and Pan Y. Minerva: A Scalable OWL Ontology Storage and Inference System. In The Semantic Web – ASWC 2006; Lecture Notes in Computer Science 4185: 429-443

Web pages

All web pages have been accessed in August 2016

[ASIA] <https://ontocrf.grupocostaisa.com/es/web/asia/home1>

[Bioportal] NCBO BioPortal <http://bioportal.bioontology.org/>

[BioTop] BioTop. <http://www2.imbi.uni-freiburg.de/ontology/biotop>

[Ca mama] <https://ontocrf.grupocostaisa.com/es/web/cancer-mama/home>

[CAPS] <https://ontocrf.grupocostaisa.com/es/web/caps/home>

[CDA]

http://www.hl7.org/implement/standards/product_brief.cfm?product_id=7

[CIMI] Clinical Information Modeling Initiative (CIMI).

<http://www.opencimi.org/>

[CMBD] <http://www.msssi.gob.es/estadEstudios/estadisticas/cmbd.htm>

[D2RQ] <http://d2rq.org/>

[Datacite] <http://www.datacite.org>

[EN 13606] <http://www.en13606.org/>

[eagle-i] <https://www.eagle-i.net/>

[FMA] Foundational Model of Anatomy.

<http://sig.biostr.washington.edu/projects/fm/AboutFM.html>

[G&H] Gastroenterología y Hepatología Continuada Web Site.

<http://www.ghcontinuada.com/>

[GO] Gene Ontology Consortium. <http://geneontology.org/>

[Health Infoway] <http://www.infoway-inforoute.ca>

[HL7 RIM] <http://www.hl7.org/implement/standards/rim.cfm>

[ICD] <http://www.who.int/classifications/icd/en/>

[ICD 9 CM] <http://www.cdc.gov/nchs/icd/icd9cm.htm>

[ISO 12967]

http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=50500

[ISO 13606-1] http://www.iso.org/iso/catalogue_detail.htm?csnumber=40784

[ISO 13606-2] http://www.iso.org/iso/catalogue_detail.htm?csnumber=50119

[ISO 13606-3] http://www.iso.org/iso/catalogue_detail.htm?csnumber=50120

[ISO 13940]

http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=58102

[ISO 21090]

http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=35646

[ISO 9000]

<https://www.iso.org/obp/ui/#iso:std:iso:9000:ed-3:v1:en>

[ISO-OSI] http://www.iso.org/iso/catalogue_detail.htm?csnumber=20269

[ISO/TS 18308:2002]

http://www.iso.org/iso/catalogue_detail.htm?csnumber=33397

[Jena] <http://jena.sourceforge.net/>

[Liferay] <http://www.liferay.com/>

[LinkEHR] <http://linkehr.com/>

[LOINC] Logical Observation Identifiers Names and Codes. <https://loinc.org/>

[LRO] <https://ontocrf.grupocostaisa.com/es/web/lro/login>

[NCI] NCI Enterprise Vocabulary Services (EVS).
<http://www.cancer.gov/research/resources/terminology>

[NEHTA] <http://www.nehta.gov.au/>

[NHS CFH] <http://www.connectingforhealth.nhs.uk>

[OBO] The OBO Foundry. <http://www.obofoundry.org/>

[openEHR] openEHR. <http://www.openehr.org/>

[OWL activity] W3C Semantic Web Activity Homepage.
<http://www.w3.org/2001/sw/wiki/OWL>

[Pellet] <http://clarkparsia.com/pellet>

[Pentaho] <http://www.pentaho.com/>

[Piscina] <https://ontocrf.grupocostaisa.com/es/web/piscina/benvingut>

[Protégé] <http://protege.stanford.edu/>.

[MSSSI] Recursos de Modelado Clínico (arquetipos). In spanish.
http://www.msssi.gob.es/profesionales/hcdsns/areaRecursosSem/Rec_mod_clinico_arquetipos.htm

[RT] <https://ontocrf.grupocostaisa.com/es/web/tumores>

[SemanticHealthNet] Semantic Interoperability for Health Network (9).
<http://www.semantichealthnet.eu/>

[SCOPE] http://cordis.europa.eu/project/rcn/78332_en.html

[SCT] SNOMED-CT. Systematized nomenclature of medicine-clinical terms.
<http://www.ihtsdo.org/snomed-ct>

[UMLS] Unified Medical Language System.

<https://www.nlm.nih.gov/research/umls/>

[VALID] <https://www.crmvf.com/>

[WAHA] <https://ontocrf.grupocostaisa.com/es/web/waha/home>

