International Journal of Medical, Medicine and Health Sciences
ISSN: 2517-9969
Vol:4, No:8, 2010

A Data Warehouse System to Help Assist Breast Cancer Screening in Diagnosis, Education and Research

Souâd Demigha

***Abstract*—**Early detection of breast cancer is considered as a major public health issue. Breast cancer screening is not generalized to the entire population due to a lack of resources, staff and appropriate tools. Systematic screening can result in a volume of data which can not be managed by present computer architecture, either in terms of storage capabilities or in terms of exploitation tools. We propose in this paper to design and develop a data warehouse system in radiology-senology (DWRS). The aim of such a system is on one hand, to support this important volume of information providing from multiple sources of data and images and for the other hand, to help assist breast cancer screening in diagnosis, education and research.

***Keywords*—**Breast cancer screening, data warehouse, diagnosis, education, research.

I. Introduction

RESEARCH on breast cancer proved that the death of breast cancer is reduced significantly if it is precociously detected and treated. Studies agree that *the mammography* with a high quality and a well interpretation, represents the more sensible and specific technique of visualization and diagnosis of breast pathologies and allows for reducing mortality [1].

To address this problem, it is necessary to create the adequate conditions allowing for the installation of mass detection campaigns, i.e. involving the maximum of women at risk. Detection is carried out starting from the analysis of breast images, primarily mammograms but also echographic images or MRI, coupled with the exploitation of information derived from the patient's history, from punctures, etc. Therefore, the clinician grounds his/her diagnosis on the result of image analysis procedures and on the synthesis of various types of information. It requires a significant amount of knowledge and know-how, which can be acquired only through a long practice [2]. At this moment, breast cancer screening is not generalized to the entire population due to a lack of resources, staff and appropriate tools. Unfortunately, systematic screening can result in a volume of data which can not be managed by present computer architecture, either in terms of storage capabilities or in terms of exploitation tools.

To overcome these deficiencies, we propose in this paper to design and develop a data warehouse system in radiology-senology (DWRS) for exploring and analysing the large amounts of data collected during the screening operation. The data warehouse system would facilitate diagnosis, education and research in order to improve the screening and thus, to reduce death per cancer and improve patient follow-up.

Clinical data are stored in proprietary commercial information systems, such as Hospital Information Systems (HIS), Radiological Information Systems (RIS) and Picture Archiving and Communication Systems (PACS) [3], intended to archive and retrieve patient records, but these systems have strict operation performance requirements and provide a little support for ad hoc queries or data analysis.

Analytical information processing often known as DSS processing (Decision Support Systems) is processing that serves the needs of management in the decision-making process. Analytical processing looks across broad vistas of data to detect trends. Instead of looking, at one or two records of data (as is the case in operational processing), with the DSS analyst does analytical processing, many records are accessed. It is rare for the DSS analyst to update data.

In operational systems, data is constantly being updated at the individual record level. In analytical processing, records are constantly being accessed, and their contents are gathered for analysis, but little or no alteration of individual records occurs. Combining Decision Support Systems and operational systems within a single information system is very difficult. Decision Support Data usually needs to be collected from a variety of operational systems (often disparate systems) and kept in a centralized data store resided on a separate platform. This corresponds to principle functioning of data warehouses systems. As Inmon [4], *data warehousing* is *subject-oriented integrated, time-variant, non volatile* collection of data in support of management decision-making. The meanings of the key terms are defined bellow:

- *Subject-oriented:* organization of data in a warehouse around the key-subjects (or high level entities) of the enterprise. For instance, patients, students and products;
- *Integrated:* the data is assumed to be using consistent naming conventions, formats, encoding, structures and related characteristics for sharing and usability;

S.Demigha is Lecturer/Researcher in Computer Science at the Department of Signal Processing and Information of EPMI: Graduate School of Electrical Engineering and Industrial Management, 13, bd de l'Hautil 95092 Cergy-Pontoise Cedex, France (corresponding author to provide phone: (33 1) 30 75 69 95; fax: (33 1) 30 75 60 41; e-mail: s.demigha@ epmi.fr).

- *Time-variant:* data contain a time dimension so that they can be used for historical purposes;
- *Non volatile:* data are refreshed from operational data and can not be updated by users.

According to [5] this principle is supported by the following factors:

- A data warehouse centralizes data (at least logically) that are scattered through out disparate operational systems and makes them readily available for decision support applications;
- A properly designed data warehouse adds value to data by improving their data quality and consistency;
- A separate data warehouse eliminates much of the contention for resources that results when informational applications are confounded with operational processes.

In most organizations, the need for data warehousing results on two major factors: a business requirement that needs to integrate company-wide of quality information and separation of informational (historical) from operational data.

For any clinical organization today, it is essential to separate operational from informational data by creating a data warehouse. Patient data is dispersed throughout the medical enterprise, being stored in proprietary systems with incompatible architectures. It is currently very difficult for radiologists to identify patient cohorts for teaching or research without enlisting the help of systems programmers who can access the data from these systems. Open-architecture data warehousing is an emerging information technology that could greatly improve access to patient data, but it has been developed in few hospital settings with little or no focus on the specific needs of radiology.

Our goal is to build a data warehouse system linking radiology reports to patient demographics, diagnosis, information, anatomo-pathology data and learning utilization patterns. The data warehouse system aims both at using the ontology (that I had developed in a previous work) [6], fitted to the description of the radiologic-senologic knowledge and on the case model (that I had developed also previously) structured as cases (with the case-based reasoning approach) in order to solve new cases by reusing them [7].

The DWRS provides:

- An ontology fitted to the description of the radiologic-senologic knowledge;
- A reuse data model (case model) of radiologic senologic images with the case-based reasoning;
- A reuse tool for diagnosis, education and research.

The paper is organized as follows:

- Section 2 presents origins of data warehouse systems;
- Section 3 presents the process of data warehouse design illustrated by a case-study in the radiology-senology domain;
- Section 4 describes the data warehouse and data models;
- Section 5 is the conclusion with further research works in progress.

Section 2 presents origins and related-work in data warehousing systems in radiology domains.

II. ORIGINS AND RELATED-WORK

The origins of data warehousing and decision support systems (DSS) processing hark back to the very early days of computers and information systems. It is interesting what DSS processing developed out of a long and complex evolution of information technology [4]. This evolution continues today.

Developing requirements for a multimedia data warehouse are different from developing requirements for data warehousing systems used in other business organizations or industries. Data warehousing did not find its way easily and readily into medicine and healthcare contrarily to other domains of industries [8]. This difference is in the form that information takes. In the business world, a transaction is repetitive and is dominated by numerical data. Contrarily, in healthcare there is no such set of repetitive transaction activity.

Data warehousing and mining techniques have rarely been applied to health care. Some researchers have initiated data warehouse and data mining projects: [9] have developed a data mining project at the Duke University Medical Center using an extensive clinical database of obstetrical patients to identify factors that contribute to perinatal outcomes. [10] have developed an Integrated Data Warehouse for radiology. Its aims were to built a data warehouse linking radiology reports to patient demographics, diagnosis, information and pathology data. [11] have developed a digital mammography data warehouse. It enables users to explore the warehouse for various analysis and decision support purposes. [3] conceived and developed a multimodality image data warehouse framework. This project applies industrial standards and methods to development of an image data warehouse framework for multimedia management, data analysis, research and access services.

These systems have some similarities with our objectives and approach our research field.

Section 3 presents the process data warehouse design adopted to our case and illustrated by a case-study in the radiology-senology domain.

III. THE PROCESS DATA WAREHOUSE DESIGN IN RADIOLOGY-SENOLOGY: A CASE-STUDY

This section describes our experience in the development of a data warehouse system in the radiology-senology domain.

A. Requirement Engineering Process

The knowledge acquisition process to build our ontology was guided by requirements with the “Crews-L'Ecritoire” approach developed in our research team in previous works [12]. It couples notions of goals and scenarios to discover knowledge. With respect to its contribution, this research has produced a step of concept extraction and described their relationships from scenarios. This approach has efficiently guided the construction of the ontology in the radiology-

senology domain. The orientation "goal" advocated is relevant since physicians and their requirements are well within the core of the process.

The "Crews-l'Ecritoire" approach is based on the ''Requirement Engineering'' concept and helps understanding users needs using a semi-automatic analysis of textual scenarios, i.e. scenarios written in natural language. Moreover, Crews permits strong control and verification of the extraction process. Starting from a high-level problem statement, it guides the discovery of a complete hierarchy of goals illustrated by scenarios in a top-down manner. The approach is based on a set of guidelines to guide linguistic analysis and verification of scenarios written in natural language. Use of natural language allows radiologists to understand scenario meaning without having expertise in Crews approach and use.

B. The MAP PROCESS: our approach

This approach is based on the case-based reasoning (CBR) and on an ontological representation of radiologic-senologic domain. The *MAP PROCESS* is a multi-step/multi-algorithm process, which permits to retrieve similar cases in various modes. As a process meta-model, it enables, thanks to the directives, a fast and simple access to knowledge. This approach relies on the formal description of the process in an intentional manner [13].

The explicit knowledge is stored on an ontological model [6], (see Fig.1), with the appropriate relations and dependencies. Experts' experience is represented as knowledge; both *product knowledge (mammographies and associated diagnoses…)* and *process knowledge (heuristics)* are considered. While the product is the result to be achieved, the process is the way the result is achieved.

The case-based reasoning is adopted to represent the experience of expert radiologists-senologists as cases [7], [14]. The case is a patient at different intervals of treatment (time). A case may comprise several successive senologic episodes. This allows to obtain an object-oriented model with the UML formalism (Unified Modelling Language) structured as cases and modelled with the MAP, a "reuse" methodology. The proposed process is on one hand, based on eliciting requirements and on the other hand, on the design of the Data Warehouse model.

Data warehouse systems aim to support decision making. Our requirement engineering process includes business process requirements (ontology), requirements from strategic decision processes (case-based reasoning) and operational data model. We combine data warehouse requirements to the data warehouse model. We have used the requirements elicitation (l'Ecritoire approach), to translate requirements in the form of business process goal. Goals and scenarios are mapped into an object-oriented data warehouse model and translate it into a dimensional data warehouse model.

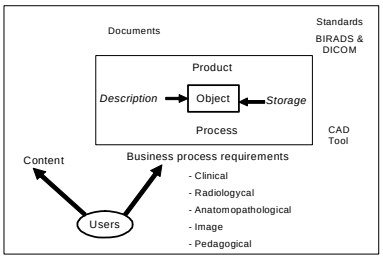

Fig. 1 The ontological representation of the DWRS

Decision analysis allows the way of guiding requirements elicitation in the data warehouse context. It is used to analyze the decisions made in business processes and in strategic decisional processes. The model used here is adopted from the *MAP PROCESS meta-model* which has been developed by the research team of the research department of computer science of the Sorbonne University [15]. Once the requirements are elicited, they are used to develop the data warehouse model (see Fig.2).

Once the new case (target case) is described by the junior radiologist-senologist, it is stored by the system. Following this, the new case is matched with all the other cases of the case base (old cases). This matching allows to find one or several relevant cases that help resolve the new problem of the trainee. A case includes one or several educational objectives. An educational objective is associated to a strategy that the expert radiologist-senologist suggests in the training of the junior-radiologist. It is one among the strategies of the MAP allowing to guide the junior-radiologist in a flexible and a supple manner. The matching is carried out via a matching algorithm.

The data warehouse will allow us on one hand, storing all the knowledge about their origins and their coming and on the other hand, with the data mining to integrate the knowledge base built with the case-based reasoning approach into the production base. The design of a case-based reasoning system includes four processes: *retrieving* the most similar cases, *reusing* the information and knowledge in the case to solve the problem, *revising* the proposed solution and *retaining* the parts of the experience useful for future problem solving. At the *Data retrieve stage*, the intervention of data mining methods can serve multiple causes. The *data cleaning* which is a part of a data mining will provide a source of "clean" information to the CBR system which will be built on accurate information. The application of data mining at this phase allows the identification of significant elements constituting a context. The CBR system must take into consideration the case and its context. Thus, the data mining will find necessary elements of the context for a given case. At the *reuse phase*, the data mining allows at once what can be done at the phases of *REtrieve and Indexing* thanks to *a decision tree, data mining techniques* allow increasing the "case retrieve" and improve answers of the CBR system when it is face to a new case. At *the Retaining phase* the data mining facilitates the

work to the developer. Figure 2 illustrates the data warehouse process with the CBR approach.

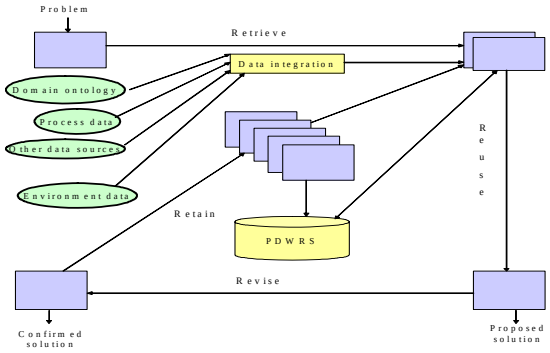

Fig. 2 The CBR Data Warehouse Process

For multimedia data mining, storage and retrieval techniques need to be integrated with standard data mining methods. Promising approaches include the construction of multimedia data cubes, the extraction of multiple features from multimedia data and similarity-based pattern matching.

C. Description of the System

As described in the previous section, on our image data warehousing system, the "Crews-l'Ecritoire" process was followed to create a logical and iterative model of the software solution. This process transforms real-word concepts such as medical images, clinical records, radiological interpretation, radiologists, anatomo-pathologists and patients and transforms them into a logical model. Requirements chunks are created to summarize operational scenarios allowed us to isolate entities with their characteristics and their relationships. This ontology allowed us to obtain a conceptual model of the domain which is structured as cases using the case-based reasoning approach.

There are two basic models that are used for database design for the warehouse: *the relational model* and *the multidimensional model.*

We find two entities in the relational model: the real world concepts and the relations between these concepts. Each entity is represented by two dimensional tables with lines and rows.

The characteristics of a concept or a relation are represented by fields at the top of each row. Each line represents a particular case (an "instance") of a concept or a relation. The advantages of the relational model are its simplicity, its normalization and the availability of a query language (SQL) that is compact and powerful. But when we face complex real world problems, relational modelling becomes difficult. Multimedia data such as we encounter in medicine are very complex, and coherence and maintenance of a relational database become delicate when data have very difficult types and have a large amount of complex relations. For example, a radiologist reviewing the case of a patient with a breast cancer may have to visualize images from various modalities (X-ray, CT, MRI, echography) with spatial relationships (two adjacent slices of a breast MRI), temporal relationships (two mammograms from the same breast with an interval period of several years), or without any relations (isolated mammograms). An "image" may be seen as an heterogeneous structure that may include all or part (aggregation of various elements) (image, image series with or without spatial and / or temporal and / or intermedia relationships, textual elements describing the image, textual elements not belonging to the image but related to it such as the age or other information such as an associated pathology from the patient record, audio elements such as a radiological interpretation report,…). To represent such a diversity, a real world concept may be scattered among several tables and thus generate non normalized relations. Such non normalized relations imply the creation of external procedures to ensure the semantic coherence of the database.

The object model allows for a higher level of abstraction when representing concepts from real word. It brings solutions to the limitations of the relational model when representing complex real world concepts. It may represent all the complexity of a medical image database system without losing information and coherence. Moreover, it allows for an easy creation of new data types and a flexible scheme adapting to new emerging imaging modalities, ever changing medical knowledge and the modifications of the model. Therefore, we have chosen object modelling for the design of our data warehouse model. We choose the Unified Modelling Language (UML) to represent the object model [16], (see Fig.3).

This data warehouse model for data representation includes data information from history and physical examination, images (primary and logical features), context (conditions in which data are created: type of acquisition, imaging modality, processing, etc.), settings and connections to other hospital systems. Section 4 describes the data warehouse and data models.

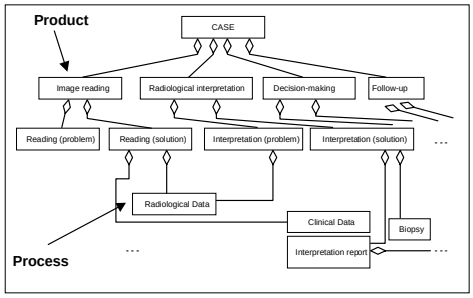

Fig. 3 Part of the dimensional model of the DWRS

IV. THE DATA WAREHOUSE AND DATA MODELS

Data warehouses and OLAP tools are based on multidimensional data model. This model views data in the form of data cube. A data cube allows data to be modeled in multiple dimensions. It is defined by dimensions and facts. In general terms, dimensions are the perspectives or entities with respect to which an organization wants to keep records. A

multidimensional data model is typically organized around a central theme. This theme is represented by a fact table. Facts are numerical measures. Think of them, as the quantities by which we want to analyze relationships between dimensions. The fact table contains the name of the facts, or measures, as well as keys to each of the related dimension tables.

A. Architecture of the Warehouse System

Figure 4 illustrates the general architecture of the data warehouse system in the radiology-senology domain. Data collected for the development of the data warehouse in radiology-senology are coming from multiple and different sources: the BI-RADS (Breast Imaging Reporting and Data System) dictionary [17], on scientific reports of the EBM (evidence-based in medicine) [18], reports and experience for radiologists-senologists of the Necker Hospital in Paris. Data acquired during data collect must be cleaned, transformed and aggregated into a single model of the data warehouse.

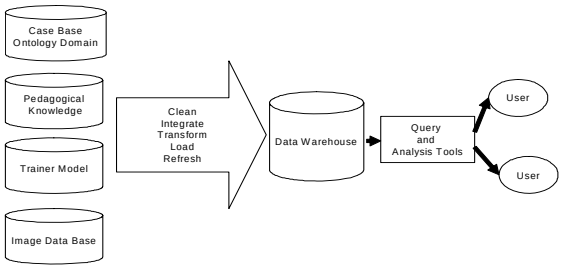

Fig. 4 General architecture of the DWRS

It is an object-oriented model with the UML language. To be compliant with commercial systems for digital mammography and CAD mammography [19], terminological systems used by the system to describe and index data must be based on DICOM [20] and BI-RADS dictionaries. The data warehouse system in radiology-senology includes four levels:

- The warehouse;
- The ontology (patient record, radiologycal and anatomo-pathologycal applications);
- Educational and research applications;
- User's interface.

The data warehouse includes two kinds of data: *expertise data* and *pedagogical data*.

Expertise knowledge

- *Clinical data*, which includes data about health patient history: screening history, current health status, and previous clinical examination;
- *Image reading data*, which consists in searching and extracting relevant information;
- *Radiological interpretation data*, which is based both on clinical data (patient's history screening, current health status, information on previous clinical examination) and radiological data (information such as defined by the BI-RADS standard);
- *Anatomo-pathological data*, which depends on the result of radiological interpretation. It grasps information about anatomo-pathological examinations such as type of procedure, reporting source, laterality, anatomo-pathology, staging and therapy. Mammogram films selected and archived during screening must be scanned and stored with the DICOM format. Digital mammogram with the DICOM format and stored will be transferred to the warehouse.

Pedagogical knowledge

Using domain knowledge (expertise of confirmed radiologists-senologists capitalized in the form of cases), the educational module develops a reasoning that allows to evaluate the trainee and to guide him/her using an educational strategy adapted to the trainee's model. The trainee's model takes into account capabilities of the trainee by proposing him/her several levels of exercises; it generates a feedback suited to each type of error. It allows understanding the origin of the error and proposes a remediation strategy; it builds a diagnosis of errors (misdiagnosis) and evaluates the trainee's work. Misdiagnoses allow more relevant and more effective interventions of the system. It is aimed at helping the trainee to use the necessary knowledge and to neglect the non-relevant one.

B. Process of Developing of the Data Warehouse

The process of developing a system of data warehouse must respect some steps. It starts by the extraction of data from the source, to transform them with very precious rules and to inject them into the target decisional system.

Data extraction

It is the process of capturing useful data from different sources. We must consider information system evolution. Data formats to be captured are heterogeneous and are being able to change in the time by evolutions of production systems as to integrate necessary supplementary data for taking into account new indicators, and to obtain a best granularity of information. At the phase of data extraction, tool usefulness is its capacity to capture data described under formats which are be to change in time and stored on systems which are be able to change too in time.

Data transformation

It consists of transforming captured data in such a way that to obtain an homogenous set of data which became comparables, additional, etc. We start by purifying captured data before transforming them. For that, we analyse data in order to identify abnormal ones which can not be introduced into the data warehouse. Data are filtered and duplications are eliminated.

Data cleaning (or data cleansing): consists of elimination of al aberrant data as data without values or with a missing value, or with a value which hasn't a particular signification. Data which can not be cleaned are ejected.

Data splitting: when data are captured from a set of databases in order to aggregate them, we are confronted with redundancies problem. These redundancies are found at a

same source or at different sources. The operation of splitting intervenes on each level required to assure the coherence and the quality of the data warehouse.

Complementary controls: classic informatics controls are used (missing required data, doubloons, coherence between various sources, cheeking of files records data, formats,…).

Data formatting and re-structuring: this operation standardizes information and prepares their introduction in the target system by converting them in the target format.

Data synchronisation: data must be synchronized, in such a way that to assure the coherence of aggregates of the data warehouse and thus, the pertinence of computed data.

Make synthesis data by computation or aggregation: the tool creates information to be injected into the data warehouse from captured data. After, it uses its meta-data referential to aggregate information thanks to defined scheme during base modelling of the data warehouse.

Data injection

This operation consists of injecting for one time collected information into the data warehouse.

Data loading

After completing the process of data preparation, data must be loaded: to transfer transformed and consolidated data into the decision database support, checking their consistency and building necessary index.

Fig. 5 The process of the DWRS

V. CONCLUSION

The present paper has presented a methodology of developing a data warehouse system in radiology-senology. It discusses the importance of defining requirements before designing and developing a data warehouse system. We have illustrated a case-study in radiology-senology by designing and developing a data warehouse system in radiology-senology. Its major objective is to facilitate access to volumes of important data useful to breast cancer screening.

The next steps to be developed would be centralised around the implementation of the system with the Java language.

REFERENCES

[1] P.A. Goumot, "Le sein son image, " *Editions Vigot*, 1993.

[2] S.Demigha and C. Rolland**. "**Training-Aided System in Senology: Methodologies and Techniques," *Conference Proceedings – SPIE Medical Imaging, PACS and Integrated Medical Information: Design and Evaluation*, vol 5033, pp339-349, San Diego, California, USA, 2003

[3] S.T.C. Wong, K.S. Hoo , R.C. Knowlton, K.D. Laxer, X. Cao, R.A. Hawkins, W.P. Dillon and L.A. Arenson, "Design and Applications of a Multimodality Image Data Warehouse Framework," *In Journal of the American Medical Informatics Association*, Vol 9, Nb 3, pp. 239-254, May/June 2002.

[4] J. W.H.Inmon, "Building the Data Warehouse," Fourth Edition, *Wiley Publishing, In Indianapolis*, Indiana, 2005.

[5] F.R. McFadden and J.A. Hoffer, "Modern Database Management: Basic Concepts of Data Warehousing," *Addison-Wesley*, New York: 1993.

[6] S. Demigha, "An Ontology Supporting the Daily Practice Requirements of Radiologists-Senologists with the Standard BI-RADS," ICEIS 2007, *Ninth International On Enterprise Information Systems, Proceedings of Information Systems Analysis and Specification*, Vol ISAS, pp 243-249, Funchal, Madeira, Portugal, June, 12-16, 2007.

[7] S. Demigha, "The retrieval Process in the SAFRS System with the Case-Based Reasoning Approach," ICEIS 2007, *Ninth International On Enterprise Information Systems, Proceedings of Artificial Intelligence and Decision Support Systems*, Vol AIDSS, pp 468-473, Funchal, Madeira, Portugal, June, 12-16, 2007.

[8] W.H. Inmon, "Data Warehousing in a Healthcare Environment," *in TDWI World Conference*, February 18-23, 2007, Las Vegas, NV.

[9] C. J. Prather, D.F. Lobach, L.K. Goodwin, R.N. J.W. Hales, M.L. Hage and W.E.Hammond, "Medical Data mining: Knowledge discovery in a Clinical Data Warehouse," *Duke University Medical Center.*

[10] D. Rubin and T. Desser, "Building an Integrated Data Warehouse for Radiology Teaching, Process Improvement, and Research," *in RSNA Radiological Society of North America*, December, 2003.

[11] H. Zhang, X.H. Cao, S.T.C. Wong, S.L. Lou and E.A. Sickles, "Developing a digital mammography data warehouse system," in *Proceedings of SPIE the International Society for Optical Engineering, Medical Imaging , PACS and Integrated Medical Information Systems: Design and Evaluation*, vol 4323, pp. 308-315, 2001.

[12] S. Demigha, C. Rolland, T.P. Baum, C. Balleyguier, B. Vincent, J. Chabriais, Y. MENU, "Crews-l'Ecritoire Analysis for the Implementation of a Medical Image Database for Mammography," *Conference Proceedings - SPIE Medical Imaging, PACS and Integrated Medical Information: Design and Evaluation* 2001, vol 4323, pp 386-396, San Diego, California, USA.

[13] S. Demigha, "The SAFRS System: a Case-Based Reasoning Training Tool for Capturing and Re-Using Knowledge," *International Journal of Computer Science and Engineering (IJCSE)*, 1:3, pp. 171-188, 2007.

[14] S. Demigha and N. Prat, "A case-based training system in radiology-senology," in *Information and Communication Technologies: from Theory to Applications (ICTTA) 04'*, Damascus, Syria, 2004.

[15] G. Grosz, C. Rolland, S. Schwer, C. Souveyet, V. Plihon, S. Si-said, C. Ben Achour, C. Gnaho, "Modelling and Engineering the Requirements Engineering Process: An Overview of the NATURE (Novel Approaches to Theory Underlying Requirements Engineering) Approach," *in Requirements Engineering Journal*, Volume 2, 1997, pp 115-131.

[16] OMG (2003), OMG, *Unified Modeling Language Specification*, March 2003, version 1.5, *http://www.omg.org.*

[17] J. Chabriais, C. Balleyguier, K. Kinkel, L. B Levey J. Stinès, and F. Thibault, "BI-RADS Illustré," *Première Edition Française basée sur la troisième Édition américaine*, 200 pages, 1998.

[18] EBM, Evidence-Based in Medicine, *http://ebem.org/index.php.*

[19] CAD, *"Computer Aided Detection for Mammography,"* Medical Systems: *www.cadxmed.com.*

[20] DICOM, "Digital Imaging and COmmunication in Medicine," *http://www.nema.org/press/020801.html.*