Hybrid Integration Approach for Heteregenous Information Sources, Applied to Medical Resources and Patients Medical Documents.

L. Cherrat[1], M. Ezziyyani[1,3], M. Ezziyyan[2], M. Bennouna[3] , M. Essaadi[1], Member, IEEE

[1] Abdel Malek Essaâdi University, Telecommunication and Information System Laboratory
[2] Abdel Malek Essaâdi University, UFR Sciences Physique d'Ingénieurs
[3] Murcie University, UFR Sciences Physique d'Ingénieurs

The goal of this research is to promote the establishment of an information system to provide to the Internet users the quality information on health and medicine related to the Patient Medial Documents (PMD). This information are extracted from several certified heterogeneous sources of the health localized dynamically. In addition, all the data of the patients and the information relating to any medical activity (operation, consultation, analysis, radiology, etc...) are stored in Data warehouse. These Data warehouse are distributed over several regions according to the geographical distribution location of the patients. Such a functional architecture will take us to deal with two integration methodologies of the heterogeneous information systems most recommended. The Datawarehous method used to store the information related to the PDM and the virtual integration method used to retrieve the health and medicine scientifics information.

In this article, we propose a new hybrid approach of integrating the heterogeneous information systems between Data warehouse and the virtual mediation. We develop a mechanism rewriting request and an algorithm adapted for updates of Data warehouse based on the mobile agents

***Index Terms*— Mediation, Data Warehouse, AXMed, hybrid integration, PMD, Health & Medical, Ontologie, rewritten query, XML.**

I. INTRODUCTION

Internet is an unprecedented free expression. It ignores national borders and cultural, and Internet-related technologies are developing at an astounding speed. Only one site may cover a wide spectrum of topics independents. If the conventional copy and audiovisual media leave physical traces, the content of the Internet is volatile.

The problem is not to find more information but to evaluate the reliability of the Internet editor as the relevance and accuracy of a document found on the Internet. The solution of this problem become more and more crucial and essential for medical information, especially with the existence of a multitude of specific medical and health sources. Indeed, in many cases, a site provides no documentation on the scientific results of a medical study, or on the available studies that support his statements concerning some treatments. In addition, several tools have been proposed to use international or national level and their actual usage has remained very limited, compared to growth in the number of Internet users concerned with health information. So, the health professionals as well as the patients deplored the difficulty in assessing the reliability of medical information and health. This requires the use of these sites by experts to exchange information and the selection and validation of information.

The main question is, how we can determine automatically and select the certified medical information resources and interrogate them through a special system in order to satisfy all medical aspects for any interrogation Internet?

Another problem that we address in this study, is the evolution of medicine that tends toward a greater acceptance of approaches to alternative medicine shows the relativity of "scientific facts'. Indeed, medical information evolves therefore constantly clinical content must include all new updated information on the date of the last days.

The goal of this research is to promote the establishment of an information system to provide to the Internet users the quality information on health and medicine related to the Patient Medial Documents (PMD). This information are extracted from several certified heterogeneous sources of the health localized dynamically. In addition, all the data of the patients and the information relating to any medical activity (operation, consultation, analysis, radiology, etc...) are stored in Data warehouse. These Data warehouse are distributed over several regions according to the geographical distribution location of the patients. Such a functional architecture will take us to deal with two integration methodologies of the heterogeneous information systems most recommended. The Data warehouse method used to store the information related to the PDM and the virtual integration method used to retrieve the health and medicine Scientifics information. The basic features of such a system can be summarized in six points:

- Simplifying the web search resources used by medical treatment of a PMD, and labeled the important sites with degree of importance and quality. This selection of certification of important sites in a manner automatic using the technique of learning during the interrogation.
- Protecting the Internet users from medical information of poor quality.
- Require the operational transparency of the site so that visitors can have all the elements that allow them to assess whether they can trust the information provided or not.
- Improve the quality of medical/health information available

through the technique of integration.

- Protecting the users against medical and health information that are imprecise or without valid scientific evidence.
- Addressing one of the main issues facing the Internet: the reliability and credibility of medical and health information.

II. CONTEXT GENERAL AND SEARCH STEPS

A. Generalities

1) Integration Need

To meet the interoperability needs of a growing number of hardware and software available today on the Web site, the design and development of a system, both flexible and efficient is needed, based on the architecture Mediator/Wrappers. The aim of such system is to intercept users' queries and find data and services most appropriate to user's queries, to define the parameters, to invoke the service and to return the result the manner transparently to users via the wrappers. These do not need to know the nature, type or location data, how services are invoked, in what language they were programmed and on what operating system they are accommodated, or any other aspects of system which is not part of the interface of the services required.

To manage these problems of integration and to ensure an effective interoperability between different services on the Web and the availability of heterogeneous resources on the Web, we propose the development of a tool that facilitates the exploitation of these resources. This software component acts as a bridge between the user and a set of heterogeneous resources whose architecture is based on the XML model. There are generally two methods of integration, Virtually Integration and Data warehouse:

2) Virtually Integration (mediator)

A mediator brings together in a unifying framework for a description of an area as well as the different sources available in relation to this area.

Thus, the user does not know the format or content sources that are available concerning its area of interest. He expressed his complaint in terms of vocabulary description field. That is a task for the mediator to input the user's query and transform it into specialized applications running on the sources, according to the description that has the content and format of these sources.

Several mediation systems have been developed so far to biology, these systems vary depending on several criteria, the objective of the tool, its data model, modelling sources, the language required by the user, and the ability to refine more or less a request.

3) Data warehouse

A Data warehouse is defined as a set of data from various sources, which vary in time and non-volatile which are used in the process of decision support. It's a big database that organizes operational data, integrates and stores to facilitate interrogation complex and analysis by giving the user an overview of information.

The concept of data warehouse offers the possibility of integrated data, consolidated and historised.

The data warehouse approach is to specialize a powerful machine, often parallel, with middleware data collection and analysis tools to manage the repository of data. A data warehouse is then a set of historical data varies over time, it is made by extracting from bases or application files, organized by subjects, consolidated in a single database, managed in a storage environment, helping to decision-making in the company.

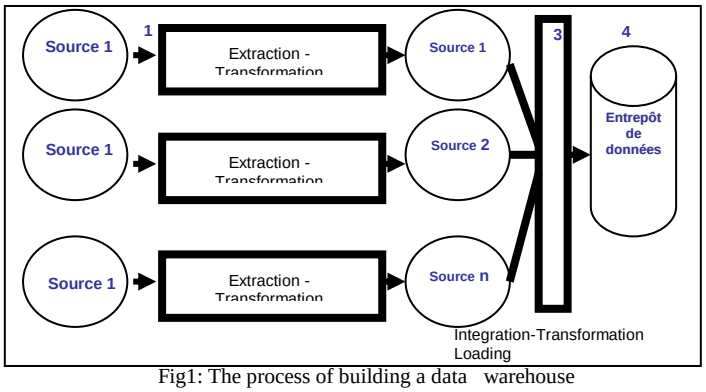


Fig1: The process of building a data warehouse

At the first level, the construction process is composed by four main phases, which are:

(1) extracting data from operational data sources, (2) the transformation of data at the structural and semantic, (3) data integration, and (4) data storage integrated into the target system. The figure below summarizes the sequence of these steps treatment.

4) AXMed Mediator

The mediator AXMed aims to offer users a convivial platform for the interrogation and integration of several heterogeneous sources collected form different servers. This system also provides the transparent tools through the integration of software resources available.

This Mediator is the result of a detailed study of the advantages and disadvantages of several existing mediators. Indeed, to meet the needs of our study and implementation of the kernel is based on technology management objects distributed around the two data models: the relational/object model and XML model. The integration approach is a mixed approach between GAV and LAV.

This last choice is justified by the fact the exploitation of the simplicity of the operation rewriting of applications by the use of GAV approach and to ensure scalability and flexibility of systems with the introduction of LAV approach. The main contribution and originality of the architecture that we propose in this system are summarized in the new methodology for defining overall pattern, the technique of managing the data warhorse using of technology agent mobile to ensure the integration of data. The generic architecture of this mediator is illustrated in the figure below:

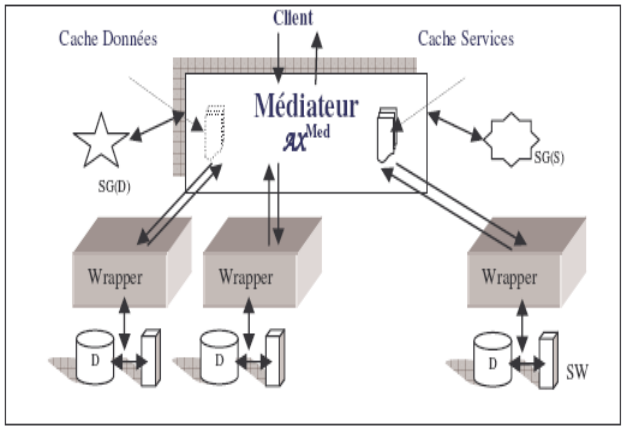


Fig 2*: generic architecture of mediator*.

III. STUDY AND MODELING

A doctor or a patient, who wishes to make finding information, is often faced with the consultation of several data sources of different origins. The absence of a standard and a single thought in the field of database design leads to the heterogeneity of these data sources. The customer is then faced with sources whose structure, semantics and language query different relative to each other.

A. Medical information and health Resources

We focus in this article on the interrogation of data sources related to genomics research cancer, this suggests a choice of integration:

- ***From clinical data (OMIM) (Online Mendelian Inheritance in Man)*** This database is a flat files, contains information on human genes and genetic disorders already mapped in the human genome. Normally, the information includes an overview of each disease, including symptoms and information on genetic mapping.

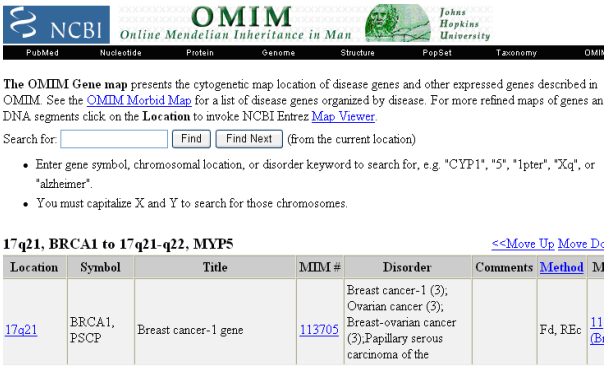
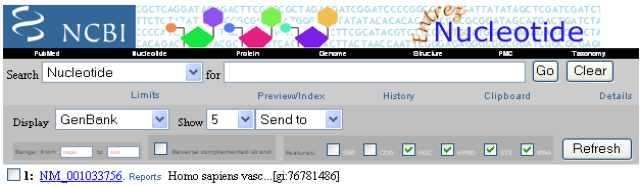


Fig 3-b*: The entry "BRCA1_HUMAN" of the OMIM source*

- ***Data on the nucleotide sequences:*** Genbank is a database containing semi-structured (XML) characterized by a collection of sequences (gene mRNA) has general information which provide information on "marital status" of the sequence : Its name, type of molecule, biological affiliation, the date of entry (LOCUS), its access number (ACCESS) as a unique identifier for registration in the bank, a brief definition and keywords to characterize, and its origin (SOURCE) and its affiliation to a biological species (ORGANISM).0

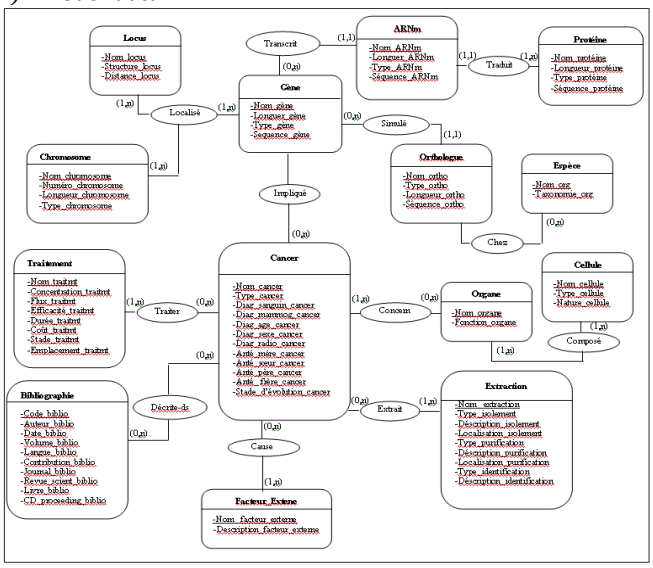



```
LOCUS       NM_001033756            3476 bp    mRNA    linear   PRI 04-OCT-2005
DEFINITION  Homo sapiens vascular endothelial growth factor (VEGF), transcript
            variant 7, mRNA.
ACCESSION   NM_001033756
VERSION     NM_001033756.1  GI:76781486
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Catarrhini;
            Hominidae; Homo.
REFERENCE   1  (bases 1 to 3476)
  AUTHORS   Fiedler,J., Leucht,F., Waltenberger,J., Dehio,C. and Brenner,R.E.
  TITLE     VEGF-A and PlGF-1 stimulate chemotactic migration of human
            mesenchymal progenitor cells
```

Fig 3-c*: The entry "BRCA1_HUMAN"of the Genbank source* [21].

As a result it has reached all the information that defines the data dictionary for the study of diseases of cancer by combining the 2 sources of genomic data. From this data dictionary and management rules (defined and established by specialist in bio-health) we tried to extract the necessary biological entities that must be handled in the study of cancer. These entities are not independent of each other and form a semantic graph whose nodes are biological entities and edges of relations between these entities.

1) Model data



B. Patient Medical Documents

The PMD will allow combining in a computerized file of information on the care that the uses were given. You can give health professionals useful information to your care, and avoid the risk of errors because they do not necessarily know what other health professionals you see or what treatment you follow. It is a guarantee of better coordination and thus better care. It is also a way to avoid duplication of redundant actions and interactions of unwanted medicines.

The medical staff will help each patient to better control its fitness. By combining in a single file the important information about your health, the health professionals will enable to adopt, whenever you request, a diagnosis better informed wherever they are.

1) Data Model

The PMD will contain information from health professionals and institutions that you have your own designated hospitals, physicians city, pharmacists, medical analysis laboratories… With a formal agreement, the case will contain including accounts and requirements doctors, the list of drugs that you have been issued, the accounts of radiology, the records of medical tests, letters of discharge from hospital. In the sequel we offer a summary of the contents of PMD:

- Strand Identification: self data in the dossier (Name, first name, date of birth, ID, etc …)
- Strand general data: data to be stored or documents to reporter.

(Personal history, medical and surgical history specialized consultations, allergy, intolerance, vaccinations)

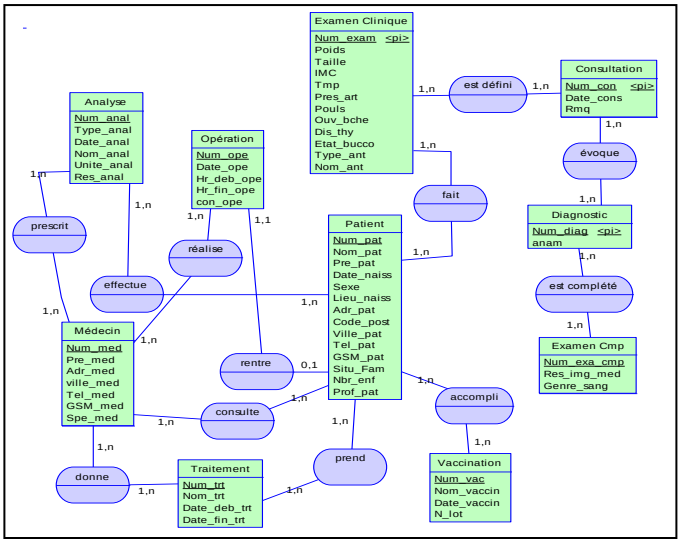
- Strand Care: documents can be deferred

(Results of biological tests, CR of diagnosis, Bilan autonomy, functional balance (physiotherapist), Conclusion of teleconsultation, CR therapeutic act, CR hospital stay, letter exit Pathology underway Dispensation Drug Follow-up care)

- Strand Prevention: documents can be deferred

(Risk Factors, CR act preventive diagnostic, therapeutic CR act referred to preventive)

- Strand Imageries: documents can be carried currently, radiological or medical imaging
- Personal space: data capture by the patient information entered at the discretion of the patient.

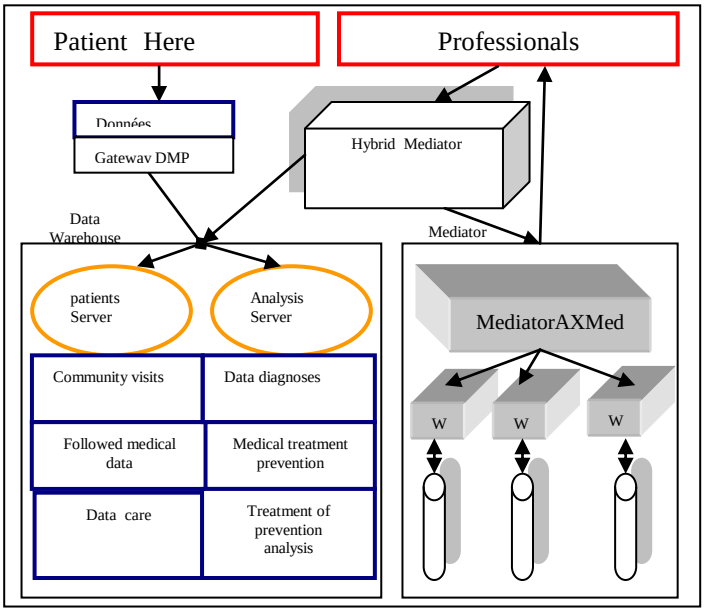


IV. FUNCTIONAL ARCHITECTURE OF INTEGRATION SYSTEM

It proposes the implementation of a system capable of uniting several sources of specific data to patients. These data can be exploited by patients, administrators, specialists, doctors. The architecture we propose is spread over several servers. Each server handles data in a specific region (analysis, consultation Radiology, Etc…). The data on each server is extract from existing health systems on the region (the concept of the data warehouse). And for the exploitation of this data on different servers, using an interface for the virtual integration using the concept of ontology. Data security remains one of the most important issues to protect servers, the level of access and transmission inter-servers. The representation of data is another problem that we address in this draft as well.

With such practical goals, the themes of scientific research in this project can address the four areas.

- Techniques Design Warehouse and algorithm updates.
- Integrating the basis of ontology algorithm and rewriting request.
- Policies and mechanisms of security for warehouses and mediators.
- Models of data representation.



V. CONCLUSION

The hybrid mediation system interface based on the AXMed mediator appears to be an answer to the problem of integration of health data distributed and heterogeneous involved in different diseases offering cost-effective results and operable. However there are still a number of improvements to the system. A possible improvement for future work would be the incorporation ontology; indeed, the creation of metadata attributes according to an ontology would assess the meaning of attributes. Thus it would be possible to link the attributes and metadata attributes according to their relationship semantics.

Another provision of the use of an ontology would be a better evaluation of results. Indeed redundancy of information, which would be an evaluation criteria, would be more easily identifiable through an ontology appropriate. Finally, the development of this system will have to go through the integration of a quantity increasingly important sources of data sources including highly specialized..

REFERENCES

[1] G. Potamiasl, A. Analyti1, D. Kafetzopoulos, M. Tsiknakis1, D. Plexousakis , P. Poirazi, M. Reczko, Y. Tollis, \Breast Cancer,

Microarrays and Biomedical Informatics: The Prognochip Project," Institute of Computer Science (ICS), FORTH, Dept. 2003, of Computer Science, University of Crete.

[2] L. Wong, \Identifying good diagnostic genes or genes groups from gene expression data by using the concept of emerging patterns," 2002, Bioinformatics 18, 725–734.

[3] R. Hesketh, "The oncogene and tumor suppressor gene facts book," 2nd ed. San Diego(CA): Academic Press, 1997, vol : 125-471

[4] R. Charles, L. Cassandra, \The science and technology behind the Human Genome chapter Finding Genes and Mutations," Smith center for Advenced Biotechnology Boston 1996, UB vol. 1820, 549-550.

[5] O. Bodenreider, \The Unified Medical Language System (UMLS): integrating biomedical terminology," Nucleic Acid Res.32, 2004, database issue:D267-70.

[6] M. Ashburner, C.Ab. Ball, J.Al. Blake, D. Botstein, H. Butler, J.Me. Cherry, A.Ph. Davis, K. Dolinski, S.Sw. Dwight, J.Th. Eppig and others, "Gene Ontology: tool for the unification of biology," The Gene Ontology Consortium. Nat. Genet 2000, , 121-267.

[7] S. Davidson, J. Crabtree, B. Brunk, J. Schug, V. Tannen, C. Overton and C. Stoeckert, \K2/Kleisli and GUS: Experiments in Integrated Access to Genomic Data Sources," IBM Systems Journal, 2001, 40 (2), 512-531.

[8] A. Bairoch, R. Apweiler, C.Hu. Wugon, W.Co. Barker, B. Boeckmann, S. Ferro, E. Gasteiger and H. Huang, \The Universal Protein Resource (UniProt)," Nucleic Acids Res. , 2005. Jan 133. Database issue:D154-9.

[9] D. Buttler, M. Coleman, T. Critchlow, R. Fileto, W. Han, C. Pu, D. Rocco, and L. Xiong, \Querying multiple Bioinformatics information sources: can Semantic web research help," ACM SIGMOD, December 2002 , Record 31.

[10] A. Cali, D. Calvanese, G. Giacomo and M. Lenzerini, \On the Expressive Power of Data Integration Systems," In Proceedings of the 21st International Conference on Conceptual Modeling , 2002 ER.

[11] Z.Bi. Miled, N. Li, M. Baumgartner and Y. Liu, \A Decentralized Approach to the Integration of Life Science Web Databases," Informatica, 2003, GHJ 27(1).

[12] L. Donelson, P.Ta. Hornoch, P. Mork, C. Dolan, J. Mitchell, M. Barrier, and H. Mei, , \The BioMediator System as a Data Integration Tool To Answer Diverse Biologic Queries," Proceedings of MedInfo, September 2004, IMIA, San Francisco, CA.

[13] M. Ezziyyani , M. Bennouna, M. Essaaidi, \Conception et Développement d'un Médiateur des Systèmes d'Information Hétérogènes "AXMed: Advanced Xml Mediator"," In Proceedings of International Symposium : ICTIS'5, 2005, pp. 122-135, Tetouan, Marocco.

[14] M.Yi. Galperin, \The Molecular Biology Database Collection: update," Nucleic Acids, Jan 2005, Res. 1 33.Database issue:D5-24.

[15] E. Guérin, B. Courselaud, and O. Loreal, \Agene Expression datawarehouse specialised in the liver," The 3rd french bioinformatics conference, proceeding, JOBIM, 2002, St Malo.

[16] N. Guarino, \Semantic Matching : Formal Ontological Distinctions for information organization, extraction and integration," In M.T. Pazienza (ed.) Information Extraction: a multidisciplinary approach to an emerging Information technology, 1998, Springer vol : 139-170.

[17] T. Hernandez, S. Kambhampati, \Integration of Biologica Sources," Current Systems and Challenges Ahead, 2004, SIGMOD Record 33(3).

[18] B. Ladjel, P. Guy, N.Du. Xuan and H. Dehainsala, \Intégration de sources de données autonomes par articulation a priori d'ontologies," 2004, Thèse de doctorat, France.

[19] P. Mork, A. Halevy, and T. Hornoch., \A Model for Data Integration Systems of Biomedical Data Applied to Online Genetic Database," In Proceedings of the Symposium Of the American Medical Informatics Association, 2001.

[20] G. Marquet, A. Burgun, F. Moussouni, and E. Guerin, , \BioMeKe : an ontology-based biomedical knowledge extraction system devoted to transcriptome analysis," Stud Health Technol Inform, 2003, source: 9580- 5.

Loubna Cherrat, Graduate student, IEEE and ASTF Member received the " Maîtrise en Informatique" degree, the "Diplôme d'Etude Supérieur Aprofondi en informatique" degree, respectively, in 2004, 2006 from Abdelmalek Essaadi University in Tangier. She is still preparing this "Doctorat National" degree in Information System Engineering.

Mostafa Ezziyyani, Prof. Dr. IEEE and ASTF Member received the " Licence en Informatique" degree, the "Diplôme de Cycle Supérieur en Informatique" degree and the "Doctorat de Troisième Cycle" degree in Information System Engineering and with honors, respectively, in 1994, 1996 and 1999 from Mahmmeed V University in Rabat, Morocco. Also, he received the "Doctorat National" degree in 2006, From Abdelmalek Essaadi University" in Distributed Systems and Web Technologies. He is a professor of Computer Engineering and Information System in the Faculty of Science and Technologies of Abdelmalek Essaadi University since 1994.