IJCSI

# International Journal of Computer Science Issues

Volume 7, Issue 3, No 2, May 2010
ISSN (Online): 1694-0784
ISSN (Print): 1694-0814


www.IJCSI.org

**IJCSI proceedings are currently indexed by:**

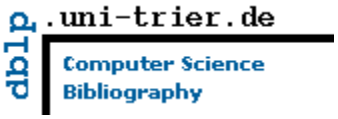
















# IJCSI Publicity Board 2010

# EDITORIAL

In this third edition of 2010, we bring forward issues from various dynamic computer science areas ranging from system performance, computer vision, artificial intelligence, software engineering, multimedia , pattern recognition, information retrieval, databases, security and networking among others.

As always we thank all our reviewers for providing constructive comments on papers sent to them for review. This helps enormously in improving the quality of papers published in this issue.

IJCSI will maintain its policy of sending print copies of the journal to all corresponding authors worldwide free of charge. Apart from availability of the full-texts from the journal website, all published papers are deposited in open-access repositories to make access easier and ensure continuous availability of its proceedings.

The transition from the 2nd issue to the 3rd one has been marked with an agreement signed between **IJCSI** and **ProQuest** and **EBSCOHOST**, two leading directories to help in the dissemination of our published papers. We believe further indexing and more dissemination will definitely lead to further citations of our authors' articles.

We are pleased to present IJCSI Volume 7, Issue 3, May 2010, split in eleven numbers (IJCSI Vol. 7, Issue 3, No. 2). The acceptance rate for this issue is 37.88%.

We wish you a happy reading!

IJCSI Editorial Board
May 2010 Issue
ISSN (Print): 1694-0814
ISSN (Online): 1694-0784

www.IJCSI.org

# IJCSI Editorial Board 2010

**Dr Vishal Goyal**
Assistant Professor
Department of Computer Science
Punjabi University
Patiala, India

**Dr Natarajan Meghanathan**
Assistant Professor
REU Program Director
Department of Computer Science
Jackson State University
Jackson, USA

**Dr Deepak Laxmi Narasimha**
Department of Software Engineering,
Faculty of Computer Science and Information Technology,
University of Malaya,
Kuala Lumpur, Malaysia

**Dr Navneet Agrawal**
Assistant Professor
Department of ECE,
College of Technology & Engineering,
MPUAT, Udaipur 313001 Rajasthan, India

**Prof N. Jaisankar**
Assistant Professor
School of Computing Sciences,
VIT University
Vellore, Tamilnadu, India

# IJCSI Reviewers Committee 2010

- Mr. Markus Schatten, University of Zagreb, Faculty of Organization and Informatics, Croatia
- Mr. Vassilis Papataxiarhis, Department of Informatics and Telecommunications, National and Kapodistrian University of Athens, Athens, Greece
- Dr Modestos Stavrakis, University of the Aegean, Greece
- Dr Fadi KHALIL, LAAS -- CNRS Laboratory, France
- Dr Dimitar Trajanov, Faculty of Electrical Engineering and Information technologies, ss. Cyril and Methodius Univesity - Skopje, Macedonia
- Dr Jinping Yuan, College of Information System and Management,National Univ. of Defense Tech., China
- Dr Alexis Lazanas, Ministry of Education, Greece
- Dr Stavroula Mougiakakou, University of Bern, ARTORG Center for Biomedical Engineering Research, Switzerland
- Dr Cyril de Runz, CReSTIC-SIC, IUT de Reims, University of Reims, France
- Mr. Pramodkumar P. Gupta, Dept of Bioinformatics, Dr D Y Patil University, India
- Dr Alireza Fereidunian, School of ECE, University of Tehran, Iran
- Mr. Fred Viezens, Otto-Von-Guericke-University Magdeburg, Germany
- Dr. Richard G. Bush, Lawrence Technological University, United States
- Dr. Ola Osunkoya, Information Security Architect, USA
- Mr. Kotsokostas N.Antonios, TEI Piraeus, Hellas
- Prof Steven Totosy de Zepetnek, U of Halle-Wittenberg & Purdue U & National Sun Yat-sen U, Germany, USA, Taiwan
- Mr. M Arif Siddiqui, Najran University, Saudi Arabia
- Ms. Ilknur Icke, The Graduate Center, City University of New York, USA
- Prof Miroslav Baca, Faculty of Organization and Informatics, University of Zagreb, Croatia
- Dr. Elvia Ruiz Beltrán, Instituto Tecnológico de Aguascalientes, Mexico
- Mr. Moustafa Banbouk, Engineer du Telecom, UAE
- Mr. Kevin P. Monaghan, Wayne State University, Detroit, Michigan, USA
- Ms. Moira Stephens, University of Sydney, Australia
- Ms. Maryam Feily, National Advanced IPv6 Centre of Excellence (NAV6) , Universiti Sains Malaysia (USM), Malaysia
- Dr. Constantine YIALOURIS, Informatics Laboratory Agricultural University of Athens, Greece
- Mrs. Angeles Abella, U. de Montreal, Canada
- Dr. Patrizio Arrigo, CNR ISMAC, italy
- Mr. Anirban Mukhopadhyay, B.P.Poddar Institute of Management & Technology, India
- Mr. Dinesh Kumar, DAV Institute of Engineering & Technology, India
- Mr. Jorge L. Hernandez-Ardieta, INDRA SISTEMAS / University Carlos III of Madrid, Spain
- Mr. AliReza Shahrestani, University of Malaya (UM), National Advanced IPv6 Centre of Excellence (NAv6), Malaysia
- Mr. Blagoj Ristevski, Faculty of Administration and Information Systems Management - Bitola, Republic of Macedonia
- Mr. Mauricio Egidio Cantão, Department of Computer Science / University of São Paulo, Brazil
- Mr. Jules Ruis, Fractal Consultancy, The Netherlands

- Mr. Mohammad Iftekhar Husain, University at Buffalo, USA
- Dr. Deepak Laxmi Narasimha, Department of Software Engineering, Faculty of Computer Science and Information Technology, University of Malaya, Malaysia
- Dr. Paola Di Maio, DMEM University of Strathclyde, UK
- Dr. Bhanu Pratap Singh, Institute of Instrumentation Engineering, Kurukshetra University Kurukshetra, India
- Mr. Sana Ullah, Inha University, South Korea
- Mr. Cornelis Pieter Pieters, Condast, The Netherlands
- Dr. Amogh Kavimandan, The MathWorks Inc., USA
- Dr. Zhinan Zhou, Samsung Telecommunications America, USA
- Mr. Alberto de Santos Sierra, Universidad Politécnica de Madrid, Spain
- Dr. Md. Atiqur Rahman Ahad, Department of Applied Physics, Electronics & Communication Engineering (APECE), University of Dhaka, Bangladesh
- Dr. Charalampos Bratsas, Lab of Medical Informatics, Medical Faculty, Aristotle University, Thessaloniki, Greece
- Ms. Alexia Dini Kounoudes, Cyprus University of Technology, Cyprus
- Mr. Anthony Gesase, University of Dar es salaam Computing Centre, Tanzania
- Dr. Jorge A. Ruiz-Vanoye, Universidad Juárez Autónoma de Tabasco, Mexico
- Dr. Alejandro Fuentes Penna, Universidad Popular Autónoma del Estado de Puebla, México
- Dr. Ocotlán Díaz-Parra, Universidad Juárez Autónoma de Tabasco, México
- Mrs. Nantia Iakovidou, Aristotle University of Thessaloniki, Greece
- Mr. Vinay Chopra, DAV Institute of Engineering & Technology, Jalandhar
- Ms. Carmen Lastres, Universidad Politécnica de Madrid - Centre for Smart Environments, Spain
- Dr. Sanja Lazarova-Molnar, United Arab Emirates University, UAE
- Mr. Srikrishna Nudurumati, Imaging & Printing Group R&D Hub, Hewlett-Packard, India
- Dr. Olivier Nocent, CReSTIC/SIC, University of Reims, France
- Mr. Burak Cizmeci, Isik University, Turkey
- Dr. Carlos Jaime Barrios Hernandez, LIG (Laboratory Of Informatics of Grenoble), France
- Mr. Md. Rabiul Islam, Rajshahi university of Engineering & Technology (RUET), Bangladesh
- Dr. LAKHOUA Mohamed Najeh, ISSAT - Laboratory of Analysis and Control of Systems, Tunisia
- Dr. Alessandro Lavacchi, Department of Chemistry - University of Firenze, Italy
- Mr. Mungwe, University of Oldenburg, Germany
- Mr. Somnath Tagore, Dr D Y Patil University, India
- Ms. Xueqin Wang, ATCS, USA
- Dr. Borislav D Dimitrov, Department of General Practice, Royal College of Surgeons in Ireland, Dublin, Ireland
- Dr. Fondjo Fotou Franklin, Langston University, USA
- Dr. Vishal Goyal, Department of Computer Science, Punjabi University, Patiala, India
- Mr. Thomas J. Clancy, ACM, United States
- Dr. Ahmed Nabih Zaki Rashed, Dr. in Electronic Engineering, Faculty of Electronic Engineering, menouf 32951, Electronics and Electrical Communication Engineering Department, Menoufia university, EGYPT, EGYPT
- Dr. Rushed Kanawati, LIPN, France
- Mr. Koteshwar Rao, K G Reddy College Of ENGG.&TECH,CHILKUR, RR DIST.,AP, India

- Mr. M. Nagesh Kumar, Department of Electronics and Communication, J.S.S. research foundation, Mysore University, Mysore-6, India
- Dr. Ibrahim Noha, Grenoble Informatics Laboratory, France
- Mr. Muhammad Yasir Qadri, University of Essex, UK
- Mr. Annadurai .P, KMCPGS, Lawspet, Pondicherry, India, (Aff. Pondicherry Univeristy, India
- Mr. E Munivel , CEDTI (Govt. of India), India
- Dr. Chitra Ganesh Desai, University of Pune, India
- Mr. Syed, Analytical Services & Materials, Inc., USA
- Dr. Mashud Kabir, Department of Computer Science, University of Tuebingen, Germany
- Mrs. Payal N. Raj, Veer South Gujarat University, India
- Mrs. Priti Maheshwary, Maulana Azad National Institute of Technology, Bhopal, India
- Mr. Mahesh Goyani, S.P. University, India, India
- Mr. Vinay Verma, Defence Avionics Research Establishment, DRDO, India
- Dr. George A. Papakostas, Democritus University of Thrace, Greece
- Mr. Abhijit Sanjiv Kulkarni, DARE, DRDO, India
- Mr. Kavi Kumar Khedo, University of Mauritius, Mauritius
- Dr. B. Sivaselvan, Indian Institute of Information Technology, Design & Manufacturing, Kancheepuram, IIT Madras Campus, India
- Dr. Partha Pratim Bhattacharya, Greater Kolkata College of Engineering and Management, West Bengal University of Technology, India
- Mr. Manish Maheshwari, Makhanlal C University of Journalism & Communication, India
- Dr. Siddhartha Kumar Khaitan, Iowa State University, USA
- Dr. Mandhapati Raju, General Motors Inc, USA
- Dr. M.Iqbal Saripan, Universiti Putra Malaysia, Malaysia
- Mr. Ahmad Shukri Mohd Noor, University Malaysia Terengganu, Malaysia
- Mr. Selvakuberan K, TATA Consultancy Services, India
- Dr. Smita Rajpal, Institute of Technology and Management, Gurgaon, India
- Mr. Rakesh Kachroo, Tata Consultancy Services, India
- Mr. Raman Kumar, National Institute of Technology, Jalandhar, Punjab., India
- Mr. Nitesh Sureja, S.P.University, India
- Dr. M. Emre Celebi, Louisiana State University, Shreveport, USA
- Dr. Aung Kyaw Oo, Defence Services Academy, Myanmar
- Mr. Sanjay P. Patel, Sankalchand Patel College of Engineering, Visnagar, Gujarat, India
- Dr. Pascal Fallavollita, Queens University, Canada
- Mr. Jitendra Agrawal, Rajiv Gandhi Technological University, Bhopal, MP, India
- Mr. Ismael Rafael Ponce Medellín, Cenidet (Centro Nacional de Investigación y Desarrollo Tecnológico), Mexico
- Mr. Supheakmungkol SARIN, Waseda University, Japan
- Mr. Shoukat Ullah, Govt. Post Graduate College Bannu, Pakistan
- Dr. Vivian Augustine, Telecom Zimbabwe, Zimbabwe
- Mrs. Mutalli Vatila, Offshore Business Philipines, Philipines
- Dr. Emanuele Goldoni, University of Pavia, Dept. of Electronics, TLC & Networking Lab, Italy
- Mr. Pankaj Kumar, SAMA, India
- Dr. Himanshu Aggarwal, Punjabi University,Patiala, India
- Dr. Vauvert Guillaume, Europages, France

- Prof Yee Ming Chen, Department of Industrial Engineering and Management, Yuan Ze University, Taiwan
- Dr. Constantino Malagón, Nebrija University, Spain
- Prof Kanwalvir Singh Dhindsa, B.B.S.B.Engg.College, Fatehgarh Sahib (Punjab), India
- Mr. Angkoon Phinyomark, Prince of Singkla University, Thailand
- Ms. Nital H. Mistry, Veer Narmad South Gujarat University, Surat, India
- Dr. M.R.Sumalatha, Anna University, India
- Mr. Somesh Kumar Dewangan, Disha Institute of Management and Technology, India
- Mr. Raman Maini, Punjabi University, Patiala(Punjab)-147002, India
- Dr. Abdelkader Outtagarts, Alcatel-Lucent Bell-Labs, France
- Prof Dr. Abdul Wahid, AKG Engg. College, Ghaziabad, India
- Mr. Prabu Mohandas, Anna University/Adhiyamaan College of Engineering, india
- Dr. Manish Kumar Jindal, Panjab University Regional Centre, Muktsar, India
- Prof Mydhili K Nair, M S Ramaiah Institute of Technnology, Bangalore, India
- Dr. C. Suresh Gnana Dhas, VelTech MultiTech Dr.Rangarajan Dr.Sagunthala Engineering College,Chennai,Tamilnadu, India
- Prof Akash Rajak, Krishna Institute of Engineering and Technology, Ghaziabad, India
- Mr. Ajay Kumar Shrivastava, Krishna Institute of Engineering & Technology, Ghaziabad, India
- Mr. Deo Prakash, SMVD University, Kakryal(J&K), India
- Dr. Vu Thanh Nguyen, University of Information Technology HoChiMinh City, VietNam
- Prof Deo Prakash, SMVD University (A Technical University open on I.I.T. Pattern) Kakryal (J&K), India
- Dr. Navneet Agrawal, Dept. of ECE, College of Technology & Engineering, MPUAT, Udaipur 313001 Rajasthan, India
- Mr. Sufal Das, Sikkim Manipal Institute of Technology, India
- Mr. Anil Kumar, Sikkim Manipal Institute of Technology, India
- Dr. B. Prasanalakshmi, King Saud University, Saudi Arabia.
- Dr. K D Verma, S.V. (P.G.) College, Aligarh, India
- Mr. Mohd Nazri Ismail, System and Networking Department, University of Kuala Lumpur (UniKL), Malaysia
- Dr. Nguyen Tuan Dang, University of Information Technology, Vietnam National University Ho Chi Minh city, Vietnam
- Dr. Abdul Aziz, University of Central Punjab, Pakistan
- Dr. P. Vasudeva Reddy, Andhra University, India
- Mrs. Savvas A. Chatzichristofis, Democritus University of Thrace, Greece
- Mr. Marcio Dorn, Federal University of Rio Grande do Sul - UFRGS Institute of Informatics, Brazil
- Mr. Luca Mazzola, University of Lugano, Switzerland
- Mr. Nadeem Mahmood, Department of Computer Science, University of Karachi, Pakistan
- Mr. Hafeez Ullah Amin, Kohat University of Science & Technology, Pakistan
- Dr. Professor Vikram Singh, Ch. Devi Lal University, Sirsa (Haryana), India
- Mr. M. Azath, Calicut/Mets School of Enginerring, India
- Dr. J. Hanumanthappa, DoS in CS, University of Mysore, India
- Dr. Shahanawaj Ahamad, Department of Computer Science, King Saud University, Saudi Arabia
- Dr. K. Duraiswamy, K. S. Rangasamy College of Technology, India
- Prof. Dr Mazlina Esa, Universiti Teknologi Malaysia, Malaysia

- Dr. P. Vasant, Power Control Optimization (Global), Malaysia
- Dr. Taner Tuncer, Firat University, Turkey
- Dr. Norrozila Sulaiman, University Malaysia Pahang, Malaysia
- Prof. S K Gupta, BCET, Guradspur, India
- Dr. Latha Parameswaran, Amrita Vishwa Vidyapeetham, India
- Mr. M. Azath, Anna University, India
- Dr. P. Suresh Varma, Adikavi Nannaya University, India
- Prof. V. N. Kamalesh, JSS Academy of Technical Education, India
- Dr. D Gunaseelan, Ibri College of Technology, Oman
- Mr. Sanjay Kumar Anand, CDAC, India
- Mr. Akshat Verma, CDAC, India
- Mrs. Fazeela Tunnisa, Najran University, Kingdom of Saudi Arabia
- Mr. Hasan Asil, Islamic Azad University Tabriz Branch (Azarshahr), Iran
- Prof. Dr Sajal Kabiraj, Fr. C Rodrigues Institute of Management Studies (Affiliated to University of Mumbai, India), India
- Mr. Syed Fawad Mustafa, GAC Center, Shandong University, China
- Dr. Natarajan Meghanathan, Jackson State University, Jackson, MS, USA
- Prof. Selvakani Kandeeban, Francis Xavier Engineering College, India
- Mr. Tohid Sedghi, Urmia University, Iran
- Dr. S. Sasikumar, PSNA College of Engg and Tech, Dindigul, India
- Dr. Anupam Shukla, Indian Institute of Information Technology and Management Gwalior, India
- Mr. Rahul Kala, Indian Institute of Inforamtion Technology and Management Gwalior, India
- Dr. A V Nikolov, National University of Lesotho, Lesotho
- Mr. Kamal Sarkar, Department of Computer Science and Engineering, Jadavpur University, India
- Dr. Mokhled S. AlTarawneh, Computer Engineering Dept., Faculty of Engineering, Mutah University, Jordan, Jordan
- Prof. Sattar J Aboud, Iraqi Council of Representatives, Iraq-Baghdad
- Dr. Prasant Kumar Pattnaik, Department of CSE, KIST, India
- Dr. Mohammed Amoon, King Saud University, Saudi Arabia
- Dr. Tsvetanka Georgieva, Department of Information Technologies, St. Cyril and St. Methodius University of Veliko Tarnovo, Bulgaria
- Dr. Eva Volna, University of Ostrava, Czech Republic
- Mr. Ujjal Marjit, University of Kalyani, West-Bengal, India
- Dr. Prasant Kumar Pattnaik, KIST,Bhubaneswar,India, India
- Dr. Guezouri Mustapha, Department of Electronics, Faculty of Electrical Engineering, University of Science and Technology (USTO), Oran, Algeria
- Mr. Maniyar Shiraz Ahmed, Najran University, Najran, Saudi Arabia
- Dr. Sreedhar Reddy, JNTU, SSIETW, Hyderabad, India
- Mr. Bala Dhandayuthapani Veerasamy, Mekelle University, Ethiopa
- Mr. Arash Habibi Lashkari, University of Malaya (UM), Malaysia
- Mr. Rajesh Prasad, LDC Institute of Technical Studies, Allahabad, India
- Ms. Habib Izadkhah, Tabriz University, Iran
- Dr. Lokesh Kumar Sharma, Chhattisgarh Swami Vivekanand Technical University Bhilai, India
- Mr. Kuldeep Yadav, IIIT Delhi, India
- Dr. Naoufel Kraiem, Institut Superieur d'Informatique, Tunisia

- Prof. Frank Ortmeier, Otto-von-Guericke-Universitaet Magdeburg, Germany
- Mr. Ashraf Aljammal, USM, Malaysia
- Mrs. Amandeep Kaur, Department of Computer Science, Punjabi University, Patiala, Punjab, India
- Mr. Babak Basharirad, University Technology of Malaysia, Malaysia
- Mr. Avinash singh, Kiet Ghaziabad, India
- Dr. Miguel Vargas-Lombardo, Technological University of Panama, Panama
- Dr. Tuncay Sevindik, Firat University, Turkey
- Ms. Pavai Kandavelu, Anna University Chennai, India
- Mr. Ravish Khichar, Global Institute of Technology, India
- Mr Aos Alaa Zaidan Ansaef, Multimedia University, Cyberjaya, Malaysia
- Dr. Awadhesh Kumar Sharma, Dept. of CSE, MMM Engg College, Gorakhpur-273010, UP, India
- Mr. Qasim Siddique, FUIEMS, Pakistan
- Dr. Le Hoang Thai, University of Science, Vietnam National University - Ho Chi Minh City, Vietnam
- Dr. Saravanan C, NIT, Durgapur, India
- Dr. Vijay Kumar Mago, DAV College, Jalandhar, India
- Dr. Do Van Nhon, University of Information Technology, Vietnam
- Mr. Georgios Kioumourtzis, University of Patras, Greece
- Mr. Amol D.Potgantwar, SITRC Nasik, India
- Mr. Lesedi Melton Masisi, Council for Scientific and Industrial Research, South Africa
- Dr. Karthik.S, Department of Computer Science & Engineering, SNS College of Technology, India
- Mr. Nafiz Imtiaz Bin Hamid, Department of Electrical and Electronic Engineering, Islamic University of Technology (IUT), Bangladesh
- Mr. Muhammad Imran Khan, Universiti Teknologi PETRONAS, Malaysia
- Dr. Abdul Kareem M. Radhi, Information Engineering - Nahrin University, Iraq
- Dr. Mohd Nazri Ismail, University of Kuala Lumpur, Malaysia
- Dr. Manuj Darbari, BBDNITM, Institute of Technology, A-649, Indira Nagar, Lucknow 226016, India
- Ms. Izerrouken, INP-IRIT, France
- Mr. Nitin Ashokrao Naik, Dept. of Computer Science, Yeshwant Mahavidyalaya, Nanded, India
- Mr. Nikhil Raj, National Institute of Technology, Kurukshetra, India
- Prof. Maher Ben Jemaa, National School of Engineers of Sfax, Tunisia
- Prof. Rajeshwar Singh, BRCM College of Engineering and Technology, Bahal Bhiwani, Haryana, India
- Mr. Gaurav Kumar, Department of Computer Applications, Chitkara Institute of Engineering and Technology, Rajpura, Punjab, India
- Mr. Ajeet Kumar Pandey, Indian Institute of Technology, Kharagpur, India
- Mr. Rajiv Phougat, IBM Corporation, USA
- Mrs. Aysha V, College of Applied Science Pattuvam affiliated with Kannur University, India
- Dr. Debotosh Bhattacharjee, Department of Computer Science and Engineering, Jadavpur University, Kolkata-700032, India
- Dr. Neelam Srivastava, Institute of engineering & Technology, Lucknow, India
- Prof. Sweta Verma, Galgotia's College of Engineering & Technology, Greater Noida, India
- Mr. Harminder Singh BIndra, MIMIT, INDIA
- Dr. Lokesh Kumar Sharma, Chhattisgarh Swami Vivekanand Technical University, Bhilai, India
- Mr. Tarun Kumar, U.P. Technical University/Radha Govinend Engg. College, India
- Mr. Tirthraj Rai, Jawahar Lal Nehru University, New Delhi, India

# TABLE OF CONTENTS

IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 3, No 2, May 2010
ISSN (Online): 1694-0784
ISSN (Print): 1694-0814


# Application Integration and Semantic Integration in Electronic Prescription Systems

**Juha Puustjärvi[1] and Leena Puustjärvi[2]**

**[1] Helsinki University of Technology**
**Espoo, Finland**

**[2] The Pharmacy of Kaivopuisto**
**Helsinki, Finland**

**Abstract**

Enterprise Application Integration (EAI) is a strategic activity and a technology set that can enable an organization to run much more efficiently, and thus provide a significant competitive advantage. Today, in the emergence of many new technologies based on Web services, there are still more approaches for EAI each approach having its limitations and opportunities. We have analyzed the limitations and opportunities of database-oriented, process-oriented, service-oriented and portal-oriented integration strategies from electronic prescription systems' point of view. It has turned out that we cannot satisfy all the requirements by any of these strategies but rather we have to develop the electronic prescription system which supports various strategies. We have also developed ontologies for semantic integration. Semantic integration has been a long-time challenge for the database community. However, by semantic integration we refer to the process of analyzing the relationships of data (i.e., their semantics) stored at communicating systems, and then using this semantics to automate the communication between computer systems. Our reason for semantic integration is the observations that within medication terms are not used in a consistent way, which have caused many serious errors in medication. In semantic integration we have used XML-based ontology languages RFD, RDF Schema and OWL.

***Keywords:*** *E-health, Electronic prescriptions, Semantic Web, Ontologies, Application Integration, Semantic Interoperability.*

## 1. Introduction

Enterprise Application Integration (EAI) concept is not new: we have been dealing with mechanisms to connect application together since we have had more than two business systems and a network to run between them. Technically, EAI is a strategic approach to binding many information systems together and supporting their ability to exchange information and leverage processes in real time. It can take two forms: organization's internal application integration and organizations external application integration. While the both forms have their own set of peculiarities they both share many common patterns.

Traditionally EAI strategies are divided into database-oriented, process-oriented, service-oriented and portal-oriented integration strategies. Today, in the emergence of many new technologies based on Web services, there are still more service-oriented approaches for EAI each approach having its limitations and opportunities.

Semantic integration has been a long-time challenge for the database community. It has received steady attention over the past decades, and has become a prominent area in information integration research. From our point of view, semantic integration is one dimension of application integration. It is the process of analyzing the relationships of data (i.e., their semantics) stored at communicating systems, and then using this semantics to automate the communication between computer systems. This process requires the development of appropriate ontologies, which are usually represented by XML-based ontology languages.

For example, during the past few years several organizations in the healthcare sector have produced standards and representation forms using XML: patient records, blood analysis and electronic prescriptions [22, 19, 12, 9, 10] are typically represented as XML-documents [1, 7] which are transferred through the SOAP-protocol [18] by accessing Web services [6, 15]. This generalization of XML-technologies sets a promising starting point for semantic integration. However, the introduction of XML itself is not enough for semantic integration but also many other more expressive XML-based technologies have to be introduced in order to achieve a seamless interoperability between the organizations within the healthcare sector.

Electronic prescription system (EPS) is an integrated system comprising of several interoperable parties. Its main function is to manage electronic prescriptions. An electronic prescription is the electronic transmission of prescriptions of pharmaceutical products from legally professionally qualified healthcare practitioners to registered pharmacies. In addition the EPS should provide a variety of services for other parties.

Choosing or developing an appropriate application integration and semantic strategy for electronic prescription systems is of prime importance as it enables health care organizations to run more efficiently, and thus provide a significant competitive advantage. In addition, the strategy should provide a suitable way for satisfying the functional requirements of all the parties of the interoperable system.

In this article we restrict ourselves on the interoperability requirements of the electronic prescription system (EPS), and describe an application integration and semantic integration strategy that aim to satisfy the requirements of the whole system. A salient feature of our approach is that it utilizes a variety of application integration strategies. At the lowest level of application integration the interaction is carried out by storing and retrieving data from the database, while at the user level interaction is hidden behind the Web services accessed by the users. In addition the EPS support process-oriented interaction through a workflow engine which in turn orchestrates Web services [15, 21], and in this sense provides a higher level integration strategy than mere Web services.

The rest of the article is organized as follows. First, in Section 2, we motivate the need of EPSs. Then, in Section 3, we describe an electronic prescription process and its functional requirements. In particular, we illustrate what kind of new facilities the deployment of the new technology should provide for electronic prescriptions processes. Then, in Section 4, we consider how three integration approaches can be adapted together in developing the EPS. Section 5, describes the XML-based standards and technologies that are used in the EPS. Then, in Section 6, we focus on semantic integration. First we give a short introduction to ontologies and then we present a simple e-prescription ontology. We also illustrate how the ontology can be utilized in prescribing medication. Finally, Section 7 concludes the article by discussing the limitations and advantages of the developed integration strategy.

## 2. Motivation for Electronic Prescriptions

The permanent trend in medication is that it increases every year. Moreover as each drug has its unique indications, cross-reactivity, complications and costs also the prescribing medication becomes still more complex every year. Fortunately, applying information and communication technology (ICT) for prescribing medication this complexity can be alleviated in many ways.

Since prescriptions are increasingly produced by the aid of ICT, it is reasonable to use ICT for transmitting prescriptions between healthcare practitioners and pharmacies, i.e., for supporting e-prescriptions (electronic prescriptions). The scope of the prescribed products varies from country to country as permitted by government authorities or health insurance carriers. The information in an electronic prescription includes for example, demographic information about the subject of care, prescribed products, dosage, amount, frequency and the details of the prescriber. The products that can be described vary from country to country as permitted by the government authorities or health insurance carriers.

The academic research of electronic prescriptions is discussed in many practitioner reports and public national plans, e.g., in [8, 2, 5, 14, 16, 10, 20]. These plans share several similar motivations and reasons for the implementation of electronic prescription systems (EPSs). These include reduction of medication errors, speeding up the prescription ordering process, better statistical data for research purposes and financial savings. The priority of these motivations varies in practitioners' reports and national plans.

## 3. The Requirements of E-Prescription Processes

We now shortly describe the e-prescription process and the data access facilities that the EPS should provide. The electronic prescription process goes as follows: first a patient visits a physician for diagnosis. In prescribing medication the physician uses the EPS. The EPS should provide versatile querying facilities on the data located in prescription holding store as well as on the data located on other healthcare systems. For example, queries about patient's previous prescriptions focus on the data stored in the prescription holding store while the queries focusing on patient's record and digital X-ray films requires the interaction with other healthcare systems.

Once the physician has constructed the prescription the EPS may automatically communicate with medical

database system in order to check (in the case of multi drug treatment) whether the drugs have mutual negative effects, and whether some of drug can be changed to cheaper ones.

After the checks and possible changes have been done the physician signs the prescription electronically. Then the prescription is encrypted and sent to an electronic prescription holding store. Basically the holding store may be centralized or distributed store. The patient will also receive the prescription in the paper form, which also includes the address of the prescription in the holding store.

The patient is usually allowed to take the prescription to any pharmacy in the country. At the pharmacy the patient gives the prescription to the pharmacist. The pharmacist will then scan the address of the prescription in the holding store by a pharmacy application, which then requests the electronic prescription from the electronic prescription holding store. After this the pharmacist will dispense the drugs to the patient and generates an electronic dispensation note. Finally they electronically sign the dispensation note and send it back to the electronic prescription holding store.

Periodically a governmental authority makes queries on the prescription holding store. These queries are statistical, focusing on the prescriptions given by certain physician or focus on the prescriptions of a patient.

The components of the EPS and the communication between the components are illustrated in Figure 1.

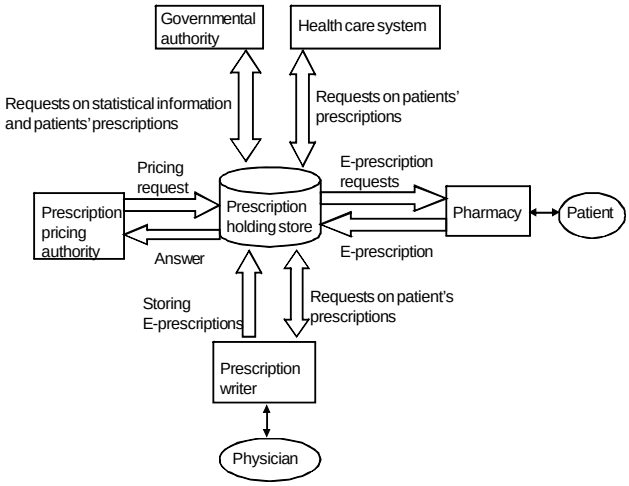



Fig. 1 The communication with the Prescription holding store.

## 4. Application Integration in Electronic Prescription Systems

Even though the approaches for the interoperation of various applications vary considerable, the principal distinction between Information-oriented, Process-oriented and Service-oriented and Portal-oriented application integration can be done [11].

- In Information-oriented approaches applications interoperate through a database.
- In Process-oriented (also called workflow–oriented [13]) approach the interoperation is controlled through a process model that binds processes and information within many systems.
- In Service-oriented interoperation applications share methods (e.g., through Web service interface) by providing the infrastructure for such method sharing.
- In Portal-oriented application integration a multitude of systems can be viewed through a single user interface, i.e., the interfaces of a multitude of systems are captured in a portal that user access by their browsers.

In our application integration strategy we use the three first of the above approaches. To illustrate this combined approach we now consider the architecture of the EPS and explain how the three integration strategies appear in it.

As illustrated in Figure 2 the EPS includes a variety of components: the users, the service requesters, the service provider and the services .

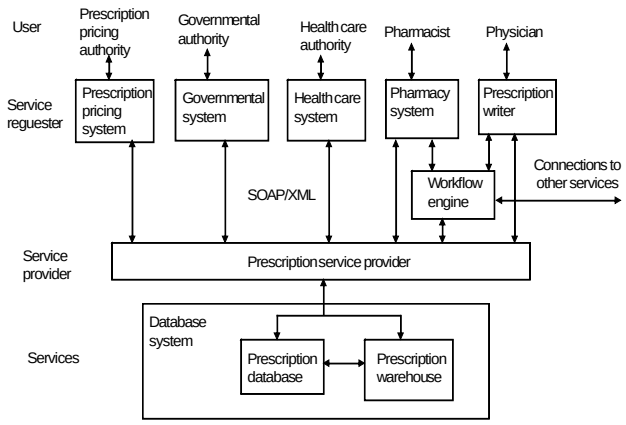



Fig. 2 The layered architecture of the EPS.

The users interact through their local applications with the Prescription service provider. Thus from users' point of view the integration strategy is service-oriented. However,

at the implementation level users interact by storing and retrieving data from the Database system which includes the Prescription database and the Prescription warehouse. In this sense the integration strategy is database-oriented. In addition, with respect to the electronic prescription process Physicians and Pharmacist interact (through Prescription writer and Pharmacy system) with the Workflow engine. The Workflow engine coordinates the execution of the prescription process which is specified by a process modeling language (explained more detail in the next section). In this sense the integration strategy is also process-oriented.

Technically the Service requester and Service provider communicates with each other using the SOAP-protocol [18]. This requires that the Service provider generates WSDL (Web Service Description Language) [17] documents which describe the interface used to invoke the services. The Service provider also provides access to the services by enabling marshalling between SOAP and the target service (Prescription database server and Prescription warehouse server). This implies interpreting the SOAP-messages and invoking the Prescription database system, and then receiving the service response and creating a SOAP response to be sent to the Service requester.

The dependencies between the Prescription database and the Prescription warehouse are the followings. Prescriptions are first stored in the Prescription database. After the prescription is dispensed, it is not necessary to store it any more in the Prescription database. Therefore the prescriptions are extracted from the Prescription database and stored to Prescription warehouse for later analysis and retrieval. However, before storing the prescriptions in the warehouse they are processed in a way that they are suitable for statistical analysis.

Basically there are three approaches to construct the data in the Prescription warehouse:

- The Prescription warehouse is periodically reconstructed from the Prescription database. During the reconstruction the Prescription database have to be done. Thus a disadvantage is the requirement of shutting done the Prescription warehouse. .

- The Prescription warehouse is changed immediately, in response to each change or small set of changes in the Prescription database. However the disadvantage of this approach is that it requires too much processing to be practical.

- The Prescription warehouse is updated periodically (e.g., each night) based on the changes that have been made to the Prescription database since the last time the warehouse was modified. This approach seems to be most suitable for the Prescription warehouse though also this choice has its disadvantages. The main disadvantage of this approach is that the changes that have been done to the Prescription database have to be calculated. Thus the update is rather complex as compared to the first approach where the Prescription warehouse is simply constructed from scratch.

## 5. The Standards and the Technologies in EPS

The architectures that use Web services require that services can be found and used. This in turn requires that the services are exactly described. The WSDL (Web Service Description Language) is an XML-based language for describing a programmatic interface to a Web service. A WSDL-description [17] includes definitions of data types, input and output message formats. For example assuming that the electronic prescription presented in Figure 3 is the input message then it is described by the XML-Schema of the prescription.

```
<Eprescription>
        <Patient>
                <Patient_name>Jack Smith </Patient name>
                <Identification> 135766677</Identification>
                <Medicine>
                        <Medicine_name>Panadol</Medicine>
                        <Disease>fewer</Disease>
                        <Quantity>15</Quantity>
                        <Refills>1</Refils>
                        <dose>One tablet three times a day</dose>
                </Medicine>
        </Patient>
</Eprescription>
```

Fig. 3 An electronic prescription in XML.P

We next consider WSCI [17] and BPEL4WS [3, 4], which are focused toward electronic prescription processes, focusing on interactions among services, as opposed to models of the services themselves (e.g., WSDL-descriptions). Thus these specifications are above the WSDL layer (Figure 4).

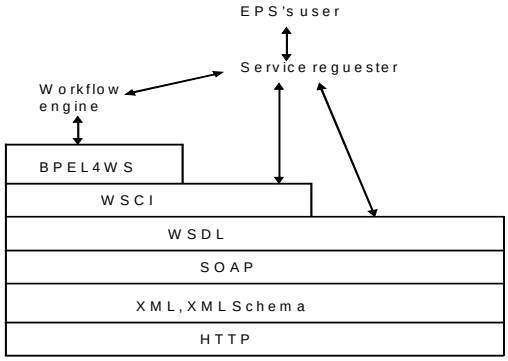



Fig. 4 The relationship of the XML-based standards.

WSCI (Web Service Choreography Interface) is an interface description language. It describes the flow of message exchanged by a Web service. By capturing the temporal and logical dependencies among the messages it characterizes the externally observable behavior of the Web service. WSCI is an enhancement of the WSDL in the sense that a WSCI specification is intended to be part of a WSDL document describing a Web service.

From electronic prescription point of view the action construct of the WSCI is very useful as it can be used to make each operation into an atomic unit of work. For example, the operations related to the retrieving an electronic prescription from the prescription database (login, prescription request, prescription response and the logout) can be made into an atomic unit.

We next illustrate the dependencies of the notion of workflow and BPE4WS. The term workflow refers to the collection of tasks organized to accomplish some business process. In out context we make the distinction between two kinds of workflows. First, we model the electronic prescription process as a workflow. Second, there are other healthcare workflows having a task which uses the services provided by the EPS. Such tasks typically retrieve patient information from the electronic prescription database. That is, retrieving information of the prescription database is a task of the workflow [13], which is implemented as a Web service request.

The tasks and their execution dependencies of the electronic prescription workflow are presented in Figure 5. In the first task the physician uses an electronic prescription writer (EPW) to construct a prescription. The electronic prescription writer (EPW) used by the physician may interact with many other health care systems in constructing the prescription. For example, the EPW may query previous prescriptions of the patient from the prescription database through the Prescription service provider. The EPW may also query patient's records from other health care systems. We assume that those health care systems provide appropriate Web services.

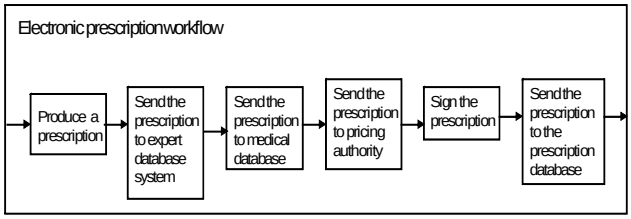



Fig. 5 An electronic prescription workflow.

Once the physician has constructed the prescription the EPW sends the prescription to the Web service of a medical expert system which checks (in the case of multi drug treatment) whether the prescribed drugs have mutual negative effects, and whether they have negative effects with other ongoing medical treatment of the patient. Then the EPW sends the prescription to the Web service of a medical database system, which checks whether the dose is appropriate. The medical database may also provide drug-specific patient education in multiple languages. It may include information about proper usage of the drug, warnings and precautions, and it can be printed to the patient. Then the EPW sends the prescription to the Web service of a pricing system, which checks whether some of the drugs can be changed to a cheaper drug. Once the checks and possible changes have been done the physician signs the prescription electronically. Then the prescription is encrypted and sent to the prescription database through the Prescription service provider.

BPEL4WS (Business Process Execution Language for Web Services) is a process modeling language for modeling the workflows which tasks are the executions of Web services. In particular, BPEL4WS description defines how multiple Web service interaction among the process's participant, called partners, are coordinated to achieve the business goal. The interactions with each partner occur through lower-level Web service interface, as might be defined in WSDL. Using BPEL4WS it is also possible to define the operations related to exception handling, including how individual or composite process components are to be compensated when exceptions and faults occur or when a partner requests an abort.

## 6. Semantic Integration in Electronic Prescription Systems

Semantic integration is one dimension of application integration. Through semantic integration we try to achieve the consistency in using terms in participating

systems. This requires the development of appropriate ontologies.

The term ontology originates from philosophy where it is used as the name of the study of the nature of existence. In the context of computer science, the commonly used definition is "An ontology is an explicit and formal specification of a conceptualization" [1]. So it is a general vocabulary of a certain domain. Essentially the used ontology must be shared and consensual terminology as it is used for information sharing and exchange. On the other hand, ontology tries to capture the meaning of a particular subject domain that corresponds to what a human being knows about that domain. It also tries to characterize that meaning in terms of concepts and their relationships.

Ontology is typically represented as classes, properties attributes and values. So they also provide a systematic way to standardize the used metadata items. Metadata items describe certain important characteristics of their target in a compact form. The metadata describing the content of a document (e.g., an electronic prescription) is commonly called semantic metadata. For example, the keywords attached to many scientific articles represent semantic metadata.

Each ontology describes a domain of discourse. It consists of a finite set of concepts and the relationship between the concepts. For example, within electronic prescription systems patient, drug, and e-prescription are typical concepts. These concepts and their relationships are graphically presented in Figure 6.

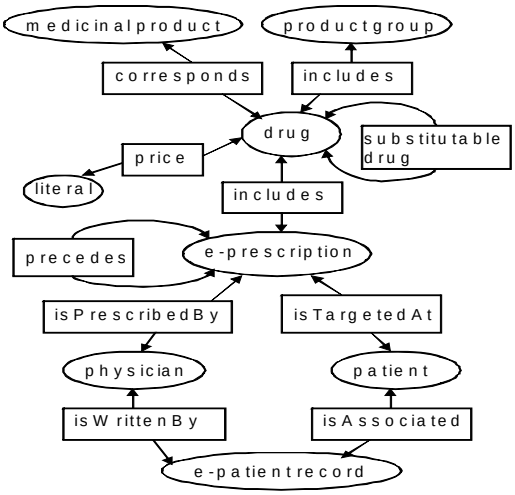



Fig. 6 An e-prescription ontology.

In the figure, ellipses are classes and boxes are properties. The ontology includes for example the following information:

- E-prescription is prescribed by a physician, and it is targeted to a patient.
- An e-prescription of a patient may precede other e-prescription of the patient, i.e., the e-prescriptions of the same patient are chained.
- Each e-prescription includes a drug.
- Each drug has a price, and it may have one or more substitutable drugs.
- Each drug corresponds a medicinal product, e.g., acetylsalicylic acid is a drug and its correspondence medicinal product is Aspirin
- Each drug belongs to a product group, e.g., aspirin belongs to painkillers.
- Each patient record is associated to a patient and it is written by a physician

The information of the ontology of Figure 6 can be utilized in many ways. For example it can be in automating the generic substitution, i.e., in changing medicinal products to cheaper medicinal products within substitutable products. It has turned out that in Finland the average price reduction of all substitutable products has been 10-15 %. In addition the automation of the generic substitution decreases the workload of the pharmacies.

In semantic integration the main problem is when we have two or more ontologies (i.e., the ontologies or conceptual scheme of various healthcare systems are not consistent), how do we find similarities between them, determine which concepts and properties represent similar notion and how to find the relationships between them. To solve these problems we capture all the concepts in the same ontology and specify the relationship of the concepts. To illustrate this, the relationships of used medicinal terms are illustrated in Figure 7. It states, for example, that the concepts tablet and pill are similarities.

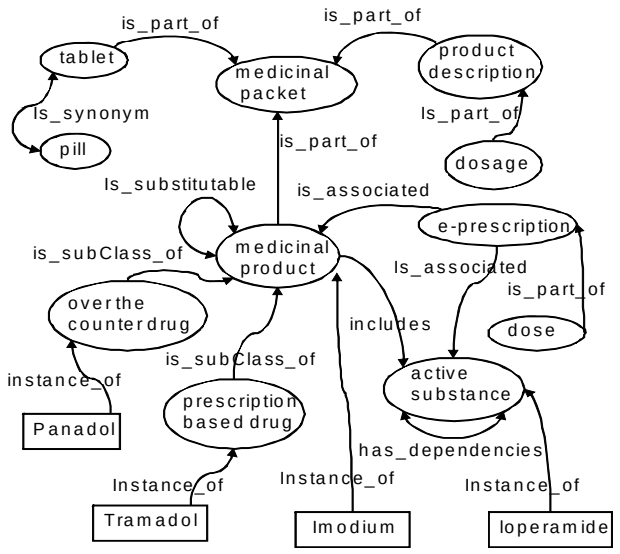



Fig. 7 A medicinal ontology.

The ontology languages we use in defining ontologies are RDF [6], RDF Schema [6] and OWL [17]. The RDF model [6] is called a triple because it has three parts: subject, predicate and object. Each triple is an RDF-statement. In order that RDF-statements can be represented and transmitted it needs syntax. The syntax has been given in XML. So an RDF-statement can be represented as an XML-document.

RDF Schema [6] provides the vocabulary for the RDF-statements. In the stack of the Semantic Web RDF-Schema is a language layered on top of RDF. It allows creating classes of data. A class is a group of things with common characteristics. For example, we have specified class e-prescription, patient and physician. Then by an RDF statement we can for example, specify that physician Jack Smith is an instance of the class physician, and by another RDF statement we can specify that prescription no. 72543 is an instance of the class e-prescription.

RDF Schema is a weak ontology language in the sense that it offers only the modelling concepts class, subclass relations, property, subproperty relation and domain and range restrictions. There are many modelling primitives that are useful in modelling documents in health care sector but are missing from RDF Schema. For example, neither we can specify that classes (e.g., physicians and patients) are not necessary disjoint nor can we build new class by set operations, e.g., class doctors is the union of the classes physicians and dentists. However, these kinds of features can be declared by the OWL (Web Ontology Language) [17].

## 7. Conclusions

Application integration is a strategic activity and a technology set that enables organizations to interoperate and to run much more efficiently, and thus provide a significant advantage. In the context of electronic prescription system the advantages arise in many ways including reduction of medication errors, speeding up the prescription ordering process, better statistical data for research purposes and financial savings.

In this article we have illustrated interoperability from electronic prescription systems point of view. In particular, we illustrated how we use different application integration strategies and XML-based semantic integration languages in developing the electronic prescription system. It has turned out that a consequence of introducing these new technologies is that it significantly changes the daily duties of the employees of the health care sector. Therefore the most challenging aspect will not be the technology but rather changing the mind-set of the employees and the training of the new technology.

The introduction of the new integration technology is also an investment. The investment on new technology includes a variety of costs including software, hardware and training costs. Training the staff on new technology is a big investment, and hence many organizations like to cut on this cost as much as possible. However, the incorrect usage and implementation of a new technology, due to lack of proper training, might turn out to be more expensive in the long run.

## References

[1] Antoniou, G. and Harmelen, F., A semantic web primer. The MIT Press. 2004.
[2] Bobbie, P., Ramisetty, S., Yussiff, A-L. and Pujari, S., Desgning an Embedded Electronic_prescription Application for Home_Based Telemedicine using OSGi Framework. http://cse.spsu.edu/pbobbie/SharedFile/NewPdfs/eRx-Final2Col.pdf, 2005.
[3] BPEL4WS – Business Process Language for Web Sevices. http://www.w.ibm.com/developersworks/webservices/library/ws-bpel/
[4] Business Process Modeling Notation (BPMN), http://www.bpmn.org/
[5]Chadwick, D., and Mundy, D., The Secure Electronic Transfer of Prescriptions. http://www.health-informatics.org/hc2004/Chadwick%20Invited%20Paper.pdf, 2004
[6] Daconta, M., Obrst, L. & Smith, K., The semantic web. Indianapolis: John Wiley & Sons. 2003.
[7] Harold, E. and Scott Means W., XML in a Nutshell. O'Reilly & Associates, 2002.

[8] Hyppönen, H., Salmivalli, L., and Suomi, R., Organizing for a National Infrastructure: The Case of the Finnish Electronic Prescription. In Proc. Of the 38th Hawaii International Conference on System Sciences.2005.
[9] Jung, F., XML-based prescription drug database helps pharmacists advise their customers. http://www.softwareag.com(xml/aplications/sanacorp.htm , 2005
[10] Keet, R.. Essential Characteristics of an Electronic Prescription Writer. Journal of Healtcare Information Management, vol 13, no 3.1999.
[11]Lithicum, D., Next Generation Application Integration. Boston: Addison-Wesley, 2004.
[12] Liu, C., Long, A., Li, Y., Tsai, K., and Kuo, H., Sharing patient care records over the World Wide Web. International journal of Medical Informatics, 61, 2001, p. 189-205.
[13] Marinescu, D., Internet-based workflow management. John Wiley & Sons, 2002.
[14] Mattocks E. 2005. Managing Medical Ontologies using OWL and an e-business Registry / Repository.
[15] Newcomer, E. Understanding Web Services. Boston: Addison-Wesley, 2003.
[16] Safran, C. and Goldberg, H.: Electronic patient records and the impact of the Internet: . International journal of Medical Informatics, 60, 2001, p. 77-83.
[17] Singh, M. and Huhns, M., Service Oriented Computing: Semantics Proceses Agents. John Wiley & Sons, 2005.
[18] SOAP – Simple Object Access Protocol. http://www.w3.org/TR/SOAP/
[19] Stalidis, G., Prenza, A. Vlachos, N., Maglavera S., Koutsouris, D. Medical support system for continuation of care based on XML web technology: International journal of Medical Informatics, 64, 2001, p. 385-400.
[20] Vesely, A., Zvarova, J., Peleska, J., Buchtela, D., and Anger, Z,. Medical guidelines presentation and comparing with Electronic Health Record, International journal of Medical Informatics, 64, 2006, p. 240-245.
[21] Web Services Activity. http://www.w3.org/2002/ws/
[22]Woolman, P. S.: XML for electronic clinical communication in Scotland. International journal of Medical Informatics, 64, 2001, p. 379-383.

**J. Puustjärvi** obtained his B.Sc. and M.Sc degree in computer science in 1985 and 1990, respectively, and his PhD degree in computer science in 1999, all from the University of Helsinki, Finland. Since 2003 he has been the professor of information society technologies at Lappeenranta University of Technology. Currently he is an adjunct professor of eBussines technologies at the Technical University of Helsinki, and an adjunct professor of computer science at the University of Helsinki. His research interests include eLearning, eHealth, eBussines, knowledge management, semantic web and databases.

**L. Puustjärvi** obtained her M.Sc degree in pharmacy in 1981, and her professional development exam in community pharmacy in 2005, both from the University of Helsinki, Finland. She has worked in several pharmacies as well as in research groups focusing on medicinal information systems. Currently she is pharmacy owner in Kaivopuisto Pharmacy in Helsinki. In addition, she participates in many medicinal researches and advises thesis focusing on medicinal information systems. Her current research interest includes electronic prescriptions, medicinal ontologies and medicinal information systems.

ISSN (Online): 1694-0784
ISSN (Print): 1694-0814

# The Vblogs: Towards a New Generation of Blogs

**Boubker Sbihi[1], Kamal Eddine El Kadiri[2] and Noura Aknin[3]**

**[1,2] Laboratoire LIROSA, Faculté des Sciences de Tétouan**
**Mhannech II, B.P : 2121 Tétouan, Morocco**

**[3] Laboratoire LaSIT, Faculté des Sciences de Tétouan**
**Mhannech II, B.P : 2121 Tétouan, Morocco**

**Abstract**

This article is part of a scientific research on the Web evolution, more precisely recent collaborative Web 2.0. It aims at suggesting an improvement of the blog, which is one of the two unique tools of publishing information at the level of Web 2.0, by trying to overcome the major problems that concern information quality, over-information and, finally, copyright management. In this context, we set the goal of answering the questions: who publishes what, when, where and what is the relevance degree of this information by adding the concept of content validation. To target this goal, we will establish a committee for blog's validation, which will be in charge of validating the information generated by users, in addition to another committee of monitoring publications on the blog, which will have as task: following the process of publication since the information creation to its deletion or final archiving. Information will not be published directly by a user on the blog but pre-published, submitted to validation in a non-official part of the website. Such ameliorations will make available to Internet users a new category of high quality information, validated and well-managed.

***Keywords:*** *Web 2.0, Vblog, validation, information quality..*

## 1. Introduction

Since its creation in 2005, the Web 2.0 or collaborative web, as advanced by Tim O'Reilly [13], is proposed as a new mode of production, communication, sharing and dissemination of information by giving the opportunity to the users to become collaborating producers of the Web content [9]. Unlike web 1.0 where most of the content was made by professionals and administrators of the Internet, Web 2.0 involves more users by generating a limitless interactivity weaving thus, social communities. This basic change is a real evolution of the web, which has undoubtedly increased the quantity of information allowing thus, the possibility of creating a collective intelligence [12]. Thanks to this vision of the web, thousands of online services on the net, among which some are for free, have emerged to replace the software acquisition and installation. According to Richard MacManus, "The web 2.0 is social, open and corresponds to new interfaces and modes of search and access. It is a ready platform to embrace educators, media, politics, communities, since practically, each has his own content "[11]. As for Devis Web 2.0 can be defined as a philosophy of social openness that aims at abolishing individual control in favor of the larger public participation "[3]. Web 2.0 is actually a series of existing technologies' use principles, suggesting a new mode of creating, publishing and sharing information on the Internet. It consists of approaching the web from different dimensions: technical, sociological and editorial. It is important to point out that the success of Web 2.0 is due to the large number of participants; for instance, in 2007, not less than 24% of Europeans published online or participated in forums [16]. Web 2.0 is characterized by the fact of possessing many tools that enable the production, the communication, sharing and dissemination of information, by involving more users and generating an infinite interactivity weaving thus, social communities. In the context of Web 2.0, users can create contents of all types, including videocasts and podcasts [6], quickly and simply by using blogs, and can comment on them, work collaboratively through wikis, introduce oneself and being introduced to others who share common interests on social networks, and be informed about news through RSS feeds. As for research, one can score, through the tags, information and share it while organizing it. Among the advantages of Web 2.0 are simplicity, flexibility, users' involvement, facility of publication, wide range of tools and the possibility of being informed about the news through RSS feeds. In the next paragraph, at the very beginning, we present the tool blog of web 2.0, then we will expose the limits of blogs namely, those related to users at first place, and then those related to information at second place in paragraph 3; thereafter, we will present the Vblog tool which is an extension of the blog tool of Web 2.0 by adding the concept of its content validation; Finally we will end the article with a conclusion where a set of perspectives is presented.

## 2. Blogs

Web 2.0 is characterized by the fact of having many tools that enable the production, communication, sharing and dissemination of information involving more users and generating an infinite interactivity weaving thus, social communities. The following table presents the most important tools of Web 2.0 and the utilities they provide:

Table 1: Utilities of Web 2.0 tools

| Tool | Utility |
|---|---|
| Blog | Regular information publication<br>Commenting information |
| Wiki | Editing content<br>Collective Intelligence |
| Social Network | Creating online communities<br>Sharing files and opinions |
| RSS feeds | Regular information monitoring |
| Tag | Improving and personalizing research |
| podcast | Sharing audio materials |
| Videocast | Sharing videos materials |
| Blog | Regular information publication<br>Commenting information |

The Web 2.0 has two types of publication tools that are blogs and wikis. Blogs are amongst the most prevailing tools on the internet, they are characterized by their simplicity, ease of use and they allow users to publish regularly information or comment on it in the context of Web 2.0. They are part of the strengths of Web 2.0 since they permit collaborative work; they are less costly or even for free and provide a space for sharing knowledge, allowing thus, a greater freedom and capacity of interaction, archiving articles and using the feature of grouping by specialty. We can define a blog or a web log (web diary) as an online diary that allows a user to publish regularly information or comment on the news about a specific topic [4]. Blogs, generally, are free to access and in case they are private, they are protected by password which replaces the vision of the instant messaging and chat. They are open to dialogue with the "Comments" function. As a content tool publishing and research and business intelligence, they allow informing users about the news in a given realm [10] and offer a vast area of knowledge sharing [2]. Like a book, a blog is identified by an IBSN (Internet Blog Serial Number), given by individuals and assigned a number in a list of over a thousand blogs. The blog can be the appointment of individuals or collective contributions of writers, journalists, consultants or members of a business, of retired experts, enthusiasts …and so on. It is a veritable collaborative communication tool [5]. Some people regards it as a journal that is available on the web, permitting an easy publication of news (articles, notes, tickets) in bloggers language) on a subject, to illustrate it with multimedia materials and share one's ideas while collecting comments on one's articles [8]. It is a space of opinions, research and creation in which one or more authors publish over time contents as texts, images, media objects and data, sorted in reverse chronological order [10]. The strengths of blogs are: the ease of use, speed and facility of publication, wide freedom, great capacity for interaction, low or even free cost, archiving of articles and the feature of specialty grouping. Blogs also have limits, the classification of limits of blogs and the limits themselves will be presented in the following paragraph.

## 3. Blogs limits

Blogs are based on user participation to create the content and form communities that gravitates around a common area of interest to share, communicate and disseminate information. However the information quality management is absent in this type of tools as long as the information may come from different types of users without considering their scientific level, their ages and specialties. In addition, no audit strategy or evaluation is taken into account. We can break down blogs' boundaries into two parts. One concerns users, while the other concerns information contained in blogs.
The types of the problems of the blogs are illustrated in the following figure

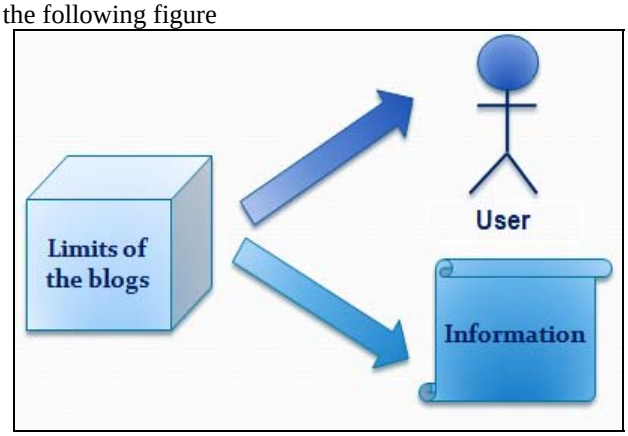



Fig. 1 The types of the problems of the blogs

### 3.1 Limits related to users

- High Rate of abandonment

The blogs are created and die quickly because of the very high dropout rate. Customers can not keep the same pace of participation all the time and quickly give up

participating in thousands of blogs per month. This is due to the large number of pages in which comments quickly increase in size with too much information while updating the blog is not regularly and promptly done.

- Few participants

Users rarely participate in the content production and satisfy themselves with simply reading, except few users who participate.

- Low participation rate

The limited number of content creators who participates in the development of blogs is characterized by a low rate of participation. This implies two types of producers, permanent producers and occasional producers.

- Heterogeneous participations

The small number of users who participate do not participate in the same manner and the same frequency; the participants are a lot when it comes to simple actions (reading, saving a bookmark, commenting), but they are less active when it comes to more complicated participation (writing, collaborating).

- The lack of motivation to participate

Blogs do not motivate users to participate regularly and regards those who produce and those who consume as identical. That is the reason why the participation rate is very low.

- Anonymous participation

Some information on blogs are anonymous, since an author can publish the information without any authentication on some sites or identify himself with a fake authentication using nicknames that do not allow to locate and show her/his true identity.

### 3.2 Limits related to Information

- Poor information quality

On blogs, everyone can create, publish, share, connect, influence, collaborate; this is a positive aspect, yet, what about the quality of what is published and to what extent is it relevant? To participate, one needs at least a minimum of knowledge and skills or may be even trainings before embarking on producing a relevant topic.

- On Information

The current mode of blogs gives a large number of pages and comments that increase in size quickly, over time, which requires readjustment from us. Everyone writes what he wants and hence there is no structure, no consistency and no convergence. It is a mixture of content coming from multiple types of users, with no classification.

- Redundant information

On blogs, you can find one ticket on several different sites. This ticket can be a translation of a text or a summary. In this context the user is lost and it becomes therefore, necessary to control the redundancy of the content by deleting redundant information and the most troublesome comments.

- Dispersed and poorly sorted information

Comments are sorted with a reverse chronological order, while one can find good comments lost among useless ones. Several new tickets of less importance come to conceal important once. No sorting according to relevance or interest is made.

- Absence of copyright

Blogs pose the problem of copyright, especially the problem of reproduction of some contents. One can even find a ticket on several sites without knowing the original source.

- Lack of security

On blogs, virtual persons are infinitely created, sometimes for propaganda purposes and publish a lot of contents on the web that violate the general principles of debating and consents.

- Very short information shelf life

The duration of the information life on blogs is very short, even if it is of great importance; as long as other important information will be created, they will replace the first important information on the main page. Information is not archived on the servers in the order of relevance, but in reverse chronological order which is a major limitation of blogs. Responsibility for the blogs amelioration is a shared responsibility; everyone must participate to address this issue and find practical and effective responses to them in order to improve them.

## 4. Vblogs

Since their emergence, blogs depend on users' participation to create them and feed them. The point here is not communicating, publishing and sharing any type of information but rather producing good information. Moreover, for the same type of user we should not regard:

- The one who produces and the one who only uses as equals.
- The one who produces a lot and the one who produces a little as equals.
- The one who produces the right information and the one who produces the wrong information as equals

That is the reason why we propose within the philosophy of Vblogs to restrict access for non-producers or require from them to pay with a virtual currency so as to access all the services and the contents.

- Identification of users

Each user must be identified by a fingerprint reader and a webcam each time he connects to the Vblog sphere. Authentication must be made on a secure site with a unique identifier on every Vblogs and will be granted once

and for ever. This new method will provide more security and less piracy and allow a user who wants to browse a clear web that only has good information; however, the user is known and can be prosecuted in case of fraud.

- Classification of information on Vblogs

Information produced on Vblogs can take several formats such as text, image, podcast and videocast. It can be divided into five classes that are represented in the following table:

Table 2: Ranking of information on Vblogs

| Symbol | Information type | Information significance |
|---|---|---|
| **G** | Good | Validated and relevant |
| **M** | Medium | Medium validated |
| **L** | Low | Submitted to corrections |
| **E** | Erroneous | Not validated |
| **C** | Opinion or comment | |

- Categories of web users

We propose creating a committee for information validation, under the responsibility of experts who will be in charge of monitoring, sorting, grading and afterwards, deleting or final archiving publications. To this end, we suggest fragmenting Vblogs' users into three groups, as represented with their roles in the following table:

Table 3: Categories of Vblogs' users

| Actor | Role |
|---|---|
| User | Reads and produces content |
| Validator | Validates what is produced |
| Expert | Publications' monitoring |

In this context, participation in a Vblog cannot be limited in reading or producing information in the form of Vtickets or Vcomments, but may also take other forms. The following figure shows the forms of participations in Vblogs by users:

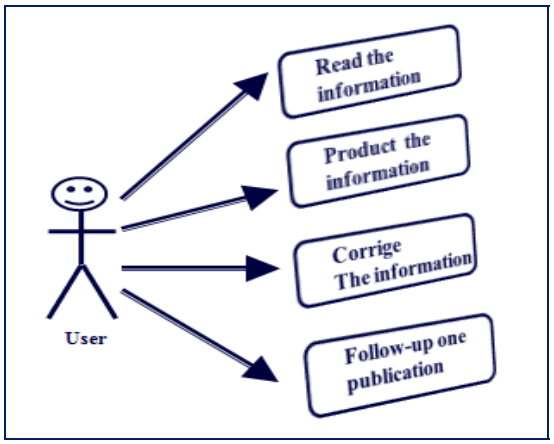



Fig. 2 Forms of participation in Vblogs.

Vblogs do not include anonymous interventions and poor quality or replicated information, but it shows the producer's identity, production date and the class of the information produced on the content.

Unlike blogs which have a single level of information, Vblogs offer four levels of quality information with one information clearly identified and sorted in order of relevance. The following figure shows the global architecture of Vblogs

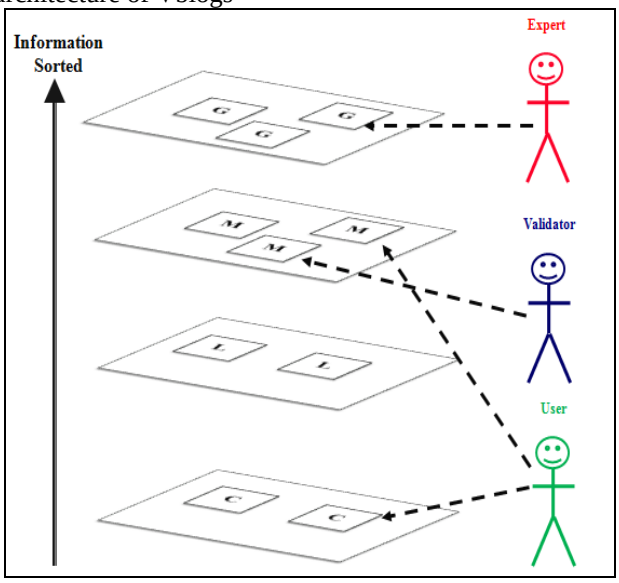



Fig. 3 Architecture of the Vblogs.

- Content validation

To validate content, an expert from the experts' community, specialized in the sphere, appoints two validators for any content submitted for publication. He publishes on the Vblog whenever both validators are in favor of the content to be published. If a validator accepts

the publication of the content while the other refuses the publication, a third validator will have the last word to decide whether the information will be published or not; However, certainly the produced information, if validated after the intervention of a third person, will be either of a medium or a low quality. In the other case, if the content was not validated it will not be published until the improvement of its quality be it a Vticket or Vcomment. A user can become a validator if s/he is recommended by two experts. In this case, the Vblogs will be organized in a hierarchical form classified, externally according to specialties and internally according to relevance, by experts who will publish only good, not redundant, non-reproduced, well sorted and validated information. They will proceed to the elimination of unnecessary information and will store the validated information in electronic archives.

- Content access

In Vblogs, our philosophy consists of limiting access for consumers and expanding it for producers and validators. Access to contents and services by types of users are represented in the following figure:

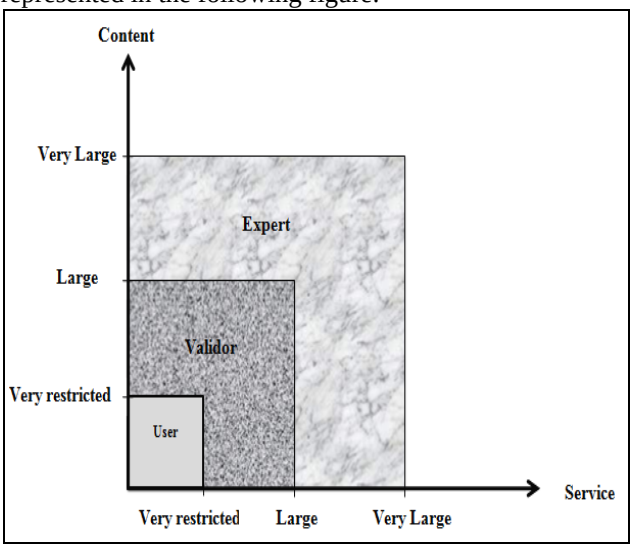



Fig. 4 Access to contents by Vblogs' user types

For instance, simple users, who do not produce the right information, will have a very limited access to Vblogs' content. Validators, who are the regular producers, will have a larger scale access by giving them the opportunity to access to the whole content made by users and the majority of Vblogs' services. Finally, experts will have the possibility to access all the content and also to additional services including secure data of Vblogs' management since they are in charge of administering it. They will have a percentage of advertisement revenues that are on the site to which they adhere. This mechanism that makes distinctions between those who participate and those who do not will urge a large population who do not participate to participate in order to have access to all documents hosted on Vblogs.

- Central Vblog

To know who published what and to avoid content duplication and to be able to manage copyrights, we propose to create a central Vblog where indexes will be stored in addition to abstracts and their producers. Before publishing validated information as Vticket or Vcomment, a check will be made at the level of the central Vblog. In case the information already exists, publication will not be allowed; otherwise, the information is accepted and will be attributed a unique number.

- Funding resources

In today's blogs, donations and advertisements are the main modes of collaborations' funding resources. In the new generation of blogs, that is to say Vblogs, the service will not be funded solely by donations or advertising; That is to say, on this part of valid and quality information of the web, in addition to the aforementioned resources we have also the contributions of users who do not produce to access various Vblogs' services.

## 4. Conclusions

Web 2.0 is the Web's new generation where the user becomes active and collaborative as opposed to Web 1.0. The problem in this version of the web is the information quality. The aim of our proposal is to answer the question about who publishes what, when and how. The information quality which is our main concern is bound to partitioning and classifying users and the produced information. In this new form, information is clearly identified, well structured, not redundant and eliminated when it is no longer useful. This will permit a better management, better reuse, and therefore a better research.

The integration of a virtual currency will urge users to participate more and take advantage from their experiences.

Many gains will be guaranteed on the levels of research time and efforts to find the right information, access will be free for producers while consumers will have to pay if they do not produce.

- Face recognition tools;
- Automatic synthesis and summary tools;
- Tools for organizing brainstorming sessions and creating collaborative projects;
- Selective dissemination of information tools.

## References

[1] Anderruthy,J (2007), Web 2.0 : Révolutions et nouveaux services d'Internet, Editions ENI, 2007.
[2] Cyril,F and Turrettini,E (2004), Blog story. Paris: Eyrolles.

[3] Devis, I. (2005). "Web 2.0 and all that ". In Internet Alchemy, [Online], Available: http://internetalchemy.org/2005/07/talis-Web-20-and-all-that
[4] Desavoye,B Moisantr,X and Le Meur,L (2005), Les blogs : nouveau média pour tous. Paris : M2 édition 2005.
[5] Ertzcheid,O (2005) , Weblogs : un nouveau paradigme pour les systèmes d'information et la diffusion de connaissances ? : Applications et cas d'usage en contexte de veille et d'intelligence économique. Colloque ISKO, France 2005.
[6] Felix, L., Stolarz, D. (2006). Hands-On Guide to Video Blogging and Podcasting: Emerging Media Tools for Business Communication. Focal Press: Massachusetts, USA.
[7] Guillaud, H.(2005), Qu'est-ce que le Web 2.0 ?, internet Actu, september.
[8] Hugh,H (2005), Blog: understanding the information what's changing your world. Nashville, 2005, ISBN 0-7852-1187-X
[9] Hussher F and al (2006), Le nouveau pouvoir des internautes, Timée-Editions, 2006.
[10] Le Meur, L and Beauvais,L (2005), Blogs pour les pro,– Paris : Dunod, 2005.
[11] MacManus, R. Collective Intelligence in Action. Pap/Onl, 2008.
[12] Musser. J. and O'Reilly.T (2006), Web 2.0 Principles and Best Practices. O'Reilly Media, Inc., 2006.
[13] O'Reilly, T. (2005). "What Is Web 2.0, Design Patterns and Business Models for the Next Generation of Software", O'Reilly Media, Inc.
[14] Sbihi, B and El kadiri,K,.(2009). "Web 2.2 : Toward classified information on the Web", International Journal of Web Applications, Vol 1, No 2, pp 102- 109.
[15] Sbihi and El kadiri,K.(2010). " Towards a participatory E-learning 2.0 :A new E-learning focused on learners and validation of the content", International Journal on Computer Science and Engineering, Vol 3, No 1.
[16] Smihily,M(2007) [Online], Eurostat, Enquête communautaire sur l'utilisation des TIC par les ménages et les particuliers, http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-QA-07-023/FR/KS-QA-07-023-FR.PDF

**Boubker Sbihi** is pHD doctor and professor of computer science at the School of Information Science in morocco. He is the responsible of Department of Information Management. He has published many articles on E-learning and Web 2.0. He is part of many boards of international journals and international conferences.

**Kamal Eddine El Kadiri** is pHD doctor and professor of computer science at Faculty of Sciences of Tétouan in Morocco. He is the Director of the ENSA School of engineers of Tetouan and the Director of LIROSA laboratory. He has published several articles on E-learning and Web 2.0. He is part of many boards of international journals and international conferences.

**Noura Aknin** is pHD doctor and professor of computer science at Faculty of Sciences of Tétouan in Morocco. She has published many articles on E-learning and Web 2.0. She is part of many boards of international journals and international conferences. She has member of the IEEE and the IEEE Computer Society

IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 3, No 2, May 2010
ISSN (Online): 1694-0784
ISSN (Print): 1694-0814


# A Multi Swarm Particle Filter for Mobile Robot Localization

**Ramazan Havangi [1], Mohammad Ali Nekoui [2] and Mohammad Teshnehlab [3]**

**[1] Faculty of Electrical Engineering, K.N. Toosi University of Technology
Tehran, Iran**

**[2] Faculty of Electrical Engineering, K.N. Toosi University of Technology
Tehran, Iran**

**[3] Faculty of Electrical Engineering, K.N. Toosi University of Technology
Tehran, Iran**

**Abstract**

Particle filter (PF) is widely used in mobile robot localization, since it is suitable for the nonlinear non-Gaussian system. Localization based on PF, However, degenerates over time. This degeneracy is due to the fact that a particle set estimating the pose of the robot looses its diversity. One of the main reasons for loosing particle diversity is sample impoverishment. It occurs when likelihood lies in the tail of the proposed distribution. In this case, most of particle weights are insignificant. To solve those problems, a novel multi swarm particle filter is presented. The multi swarm particle filter moves the samples towards region of the state space where the likelihood is significant, without allowing them to go far away from the region of significant values for the proposed distribution. The simulation results show the effectiveness of the proposed algorithm.

**Keywords**: Localization, Particle Filter, Particle Swarm Optimization (PSO)

## 1. Introduction

Mobile localization is the problem of estimating a robot's pose (location, orientation) relative to its environment. It represents an important role in the autonomy of a mobile robot. From the viewpoint of probability, the localization problem is a state estimation process of a mobile robot. Many existing approaches rely on the kalman filter (KF) for robot state estimation. But it is very difficult to be used in practice since KF can only be used in Gaussian noise and linear systems. To solve the problem of nonlinear filtering, the extended kalman filter (EKF) was proposed. The localization based on EKF was proposed in [1], [2], [3], [4], [5], [6] for the estimation of robot's pose. However, the localization based on EKF has the limitation that it doses not apply to the general non-Gaussian distribution. In order to represent non-linearity and non-Gaussian characteristics better, particle filter was proposed in [19], [20]. Particle filter outperforms the EKF for nonlinear systems and has been successfully used in robotics. In recent years, the particle filter (PF) is widely used in localization [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20]. The central idea of particle filters is to represent the posterior probability density distribution of the robot by a set of particles with associated weights. Therefore, the particle filters do not involve linearzing the models of the system and are able to cope with noises of any distribution. However, localization based on particle filter also has some drawbacks. In [19], [20], [21], [22], [23], [24], it has been noted that it degenerates over time. This degeneracy is due to the fact that particle set estimating the pose of the robot looses its diversity. One of main reasons for loosing particle diversity is sample impoverishment. It occurs when likelihood is highly peaked compared to the proposed distribution, or lies in the tail of the proposed distribution. On the other hand, PF highly relies on the number of particles to approximate the distribution density [19], [20], [21], [22], [23], [24]. Researchers have been trying to solve those problems in [21], [22], [23], and [24]. In all the aforementioned studies, the reliability of measurement plays a crucial role in the performance of the algorithm and additive noise was considered only. In this paper to solve those problems, a novel multi swarm particle filter is purposed. The multi swarm particle filter move samples towards the region of the state space where the likelihood is significant, without allowing them to go far away from the region of significant values of the proposed distribution. For this purpose, the multi swarm particle filter employs a conventional multi objective optimization approach to weight the likelihood and prior of the filter in order to alleviate the particle impoverishment problem. The minimization of the corresponding objective function is performed using the Gaussian PSO algorithm,

## 2. Kinematics Modeling Robot and its Odometery

The state of robot can be modeled as $(x, y, \theta)$ where $(x, y)$ are the Cartesian coordinates and $\theta$ is the orientation respective to the global environment. The kinematics equations for the mobile robot are in the following form [1-2] and [4]:

$$\begin{bmatrix} \dot{x} \\ \dot{y} \\ \dot{\phi} \end{bmatrix} = f(X) = \begin{bmatrix} (V + v_v)\cos(\phi + [\gamma + v_\gamma]) \\ (V + v_v)\sin(\phi + [\gamma + v_\gamma]) \\ \frac{(V + v_v)}{B}\sin(\gamma + v_\gamma) \end{bmatrix} \quad (1)$$

Where $B$ is the base line of the vehicle and $u = [V \quad \gamma]^T$ is the control input consisting of a velocity input $V$ and a steer input $\gamma$, as shown in Fig.1.

The process noise $v = [v_v \quad v_\gamma]^T$ is assumed to be applied to the control input, $v_v$ to velocity input, and $v_\gamma$ to the steer angle input. The vehicle is assumed to be equipped with a sensor (range-laser finder) that provides a measurement of range $r_i$ and bearing $\theta_i$ to an observed feature $\rho_i$ relative to the vehicle as follows:

$$\begin{bmatrix} r_i \\ \theta_i \end{bmatrix} = h(X) = \begin{bmatrix} \sqrt{(x - x_i)^2 + (y - y_i)^2} + \omega_r \\ \tan^{-1}\frac{y - y_i}{x - x_i} - \phi + \omega_\theta \end{bmatrix} \quad (2)$$

where $(x_i, y_i)$ is the position landmark in the map and $W = [\omega_r \quad \omega_\theta]^T$ relates to the observation noise.

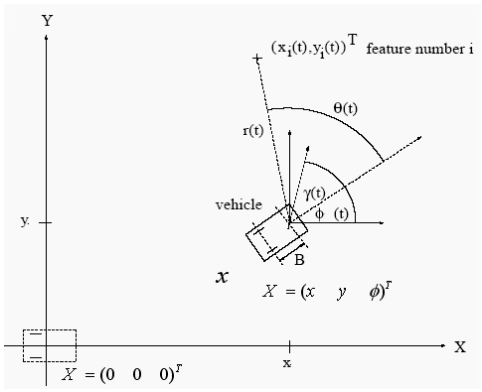



Fig.1 The Robot and Feature

## 3. Particle filter Principle

The particle filter is a special version of the Bayes filter, and is based on sequential Monte Carlo (SMC) sampling. A dynamic system represented by

$$x_k = f(x_{k-1}, \omega_k) \quad (3)$$

$$y_k = h(x_k, v_k) \quad (4)$$

is considered, where $x_k \in R^n$ is the state vector and $y_k \in R^m$ is an output vector. $f(.)$ and $g(.)$ denote the system and measurement equations, respectively. $\omega_k$ and $v_k$ are independent white-noise variables. Particle filter represents the posterior probability density function $p(x_k \mid y_{1:k})$ by a set of random samples with associated weight as follows [19], [20]:

$$S_k = \{(x_k^i, w_k^i) \mid i = 1, ..., N\} \quad (5)$$

where $x_k^i$ denotes the $i$ th particle of $S_k$, $w_k^i$ is the associated importance weight and $y_{1:k}$ denotes the measurements accumulated up to k . Then, the posterior density $p(x_k \mid y_{1:k})$ can be approximated as follows [19], [20]:

$$p(x_k \mid y_{1:k}) \approx \sum_{i=1}^{n} w_k^i \delta(x_k - x_k^i) \quad (6)$$

Where $\delta(x)$ is Dirac's delta function ( $\delta(x) = 1$ for $x = 0$ and $\delta(x) = 0$ otherwise), and $w_k^{(i)}$ is associated weight $x_k^i$ with $w_k^i > 0$, $\sum_{i=1}^{n} w_k^i = 1$. In general, it is not possible to draw samples directly from posterior $p(x_k \mid y_{1:k})$. Instead, the samples are drawn from a simpler distribution called the proposed distribution $q(x_k \mid y_{1:k})$ .The mismatch between the posterior and proposed distributions is corrected using a technique called importance sampling. Therefore, in regions where the target distribution is larger than the proposed distribution, the samples are assigned a larger weigh. Also, in regions where the target distribution is smaller than the proposed distribution the samples will be given lower weights. An example of importance sampling is shown in Fig. 2

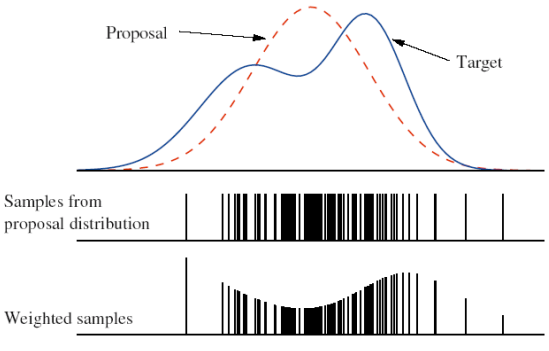



Fig2.Important Sampling

As a result, the important weight of each particle is equal to the ratio of the target posterior and the distribution proposal as follows:

$$w_k^i = \frac{\text{target distribution}}{\text{proposal distribution}} = \frac{p(x_k^i \mid y_{1:k})}{q(x_k^i \mid y_{1:k})} \quad (7)$$

The proposed $q(x_k^i \mid y_{1:k})$ can be represented by a recursive form as:

$$q(x_k^i \mid y_{1:k}) = q(x_k^i \mid x_{k-1}^i, y_{1:k}) q(x_{k-1}^i \mid y_{1:k}) \overset{Markov}{=} q(x_k^i \mid x_{k-1}^i, y_{1:k}) q(x_{k-1}^i \mid y_{1:k-1}) \quad (8)$$

Then one can obtain samples $x_k^i \sim q(x_k \mid y_{1:k})$ by augmenting each of the exiting samples $x_{k-1}^i \sim q(x_{k-1} \mid y_{1:k-1})$ with the new state $x_k^i \sim q(x_k \mid X_{k-1}, y_{1:k})$. Similarity, the posterior can also be given by a recursive form using Bayes rule as follows:

$$p(x_k^i \mid y_{1:k}) = \frac{p(y_k \mid x_k^i, y_{1:k-1}) p(x_k^i \mid y_{1:k-1})}{p(y_k \mid y_{1:k-1})}$$
$$= \frac{p(y_k \mid x_k^i, y_{1:k-1}) p(x_k^i \mid x_{k-1}^i, y_{1:k-1})}{p(y_k \mid y_{1:k-1})} p(x_{k-1}^i \mid y_{1:k-1}) \quad (9)$$
$$= \frac{p(y_k \mid x_k^i, y_{1:k-1}) p(x_k^i \mid x_{k-1}^i)}{p(y_k \mid y_{1:k-1})} p(x_{k-1}^i \mid y_{1:k-1})$$
$$\propto p(y_k \mid x_k^i, y_{1:k-1}) p(x_k^i \mid x_{k-1}^i) p(x_{k-1}^i \mid y_{1:k-1})$$

Therefore, a sequential importance weight of the $m$ th particle can be obtained as follows:

$$w_k^i \propto \frac{p(x_k^i \mid y_{1:k})}{q(x_k \mid x_{k-1}, y_{1:k-1}) q(x_{k-1} \mid y_{1:k-1})} \quad (10)$$

$$w_k^i \propto \frac{p(y_k \mid x_k^i) p(x_k^i \mid x_{k-1}^i) p(x_{k-1}^i \mid y_{1:k-1})}{q(x_k^i \mid x_{k-1}^i, y_{1:k}) q(x_{k-1}^i \mid y_{1:k-1})}$$
$$= w_{k-1}^i \frac{p(y_k \mid x_k^i) p(x_k^i \mid x_{k-1}^i)}{q(x_k^i \mid x_{k-1}^i, y_{1:k})} \quad (11)$$

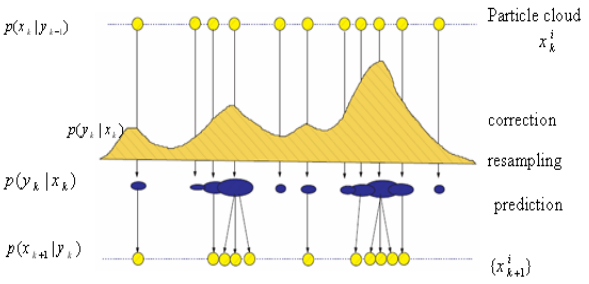



Fig.3 An illustration of generic particle filter with importance sampling and resampling.

A particle filter described above is called the sequential Importance Sampling (SIS). The SIS algorithm has a problem that it degenerates quickly over time. In practical terms this means that after a certain number of recursive steps, most particles will have negligible. Degeneracy can be reduced by using a resampling step [19], [20]. Resampling is a scheme to eliminate particles small weights and to concentrate and replace on particles with large weights. Fig.3 shows the generic particle filter with importance sampling and resampling.

## 4. Localization Based on Particle Filter

From the viewpoint of Bayesian, Mobile robot localization is basically a probability density estimation problem. In fact, Localization is estimating the posterior probability density of the robot's pose relative to a map of its environment. Assuming that the robot's pose at time $k$ is denoted by $x_k$ and measurements up to time $k$ is denoted by $Y_k$, the posterior probability distribution is as follow:

$$p(x_k \mid Y_k, m) \quad (12)$$

Where $m$ is the map of the environment which is known. The measurement data $Y_t$ comes from two different sources: motion sensors which provide data relating to the change of situation (e.g., odometer readings) and perception sensors which provide data relating to the environment (e.g., laser range scans). In other words, measurement data can be divided in two groups of data as $Y_k = \{Z_k, U_{k-1}\}$ where $Z_k = \{y_0, ..., y_k\}$ contains the range laser finder measurements and $U_{k-1} = \{u_0, ..u_{k-1}\}$ contains the odometric data. The Bayesian recursive determination of the posterior density can be computed in two steps:

1) Measurement update

$$p(x_k \mid Y_k, m) = \frac{p(y_k \mid x_k, Y_{k-1}, m) p(x_k \mid Y_{k-1}, m)}{p(y_k \mid Y_{k-1}, m)}$$
$$= \frac{p(y_k \mid x_k, m) p(x_k \mid Y_{k-1}, m)}{p(y_k \mid Y_{k-1}, m)} \quad (13)$$

where

$$p(y_k \mid Y_{k-1}, m) = \int p(y_k \mid x_k .m) p(x_k \mid Y_{k-1}, m) dx_k \quad (14)$$

2) Prediction

$$p(x_{k+1} \mid Y_k, m) = \int p(x_{k+1} \mid x_k, u_k, Y, m_k)$$
$$p(x_k \mid Y_{k-1}, m) dx_k = \int p(x_{k+1} \mid x_k, u_k, m) \quad (15)$$
$$p(x_k \mid Y_{k-1}, m) dx_k$$

The localization based on PF represents the posterior probability density function $p(x_k \mid Y_k, m)$ with $N$ weighted samples

$$p(x_k \mid Y_k, m) = \sum_{i=1}^{N} w_k^i \delta(x_k - x_k^i) \quad (16)$$

The localization algorithm of the mobile robot is realized using particle filter as following:

1. Sampling a new robot pose.

3. Calculate importance weight and normalization.
4. Normalized Wight
The normalized weights are given by:

$$w_k^i = \frac{w_k^i}{\sum_{i=1}^{N} w_k^i} \quad (17)$$

4. Resampling
In the following subsections we give details of the main steps. To alleviate the notation, the term $m$ is not included in the following expressions, $p(x_k | Y_k)$.

### 4.1 Sampling a New of Pose

The choice of importance density $q(x_k^i | x_{k-1}^i, Y_k)$ is one of the most critical issues in the design of a particle filter. Two of those critical reasons are as follows: samples are drawn from the proposed distribution, and the proposed distribution is used to evaluate important weights. The optimal importance density function minimizes the variance of the importance weights through the following equation [19], [20].

$$q_{opt}(x_k^i | x_{k-1}^i, Y_k) = p(x_k^i | x_{k-1}^i, Y_k) \quad (18)$$

However, there are some special cases where the use of the optimal importance density is possible. The most popular suboptimal choice is the transitional prior

$$q_{opt}(x_k^i | x_{k-1}^i, Y_k) = p(x_k^i | x_{k-1}^i) \quad (19)$$

In this paper, the proposed distribution in equation (19) is used due to its easy calculation. Hence, by the substitution of (19) into (11), the weight's update equation is:

$$w_k^i \propto w_{k-1}^i p(y_k | x_k^i) \quad (20)$$

### 4.2 Resampling

Sine the variance of the importance weights increases over time [21], [23], [25], resampling plays a vital role in the particle filter. In the resampling process, particles with low importance weight are eliminated and particles with high weights are multiplied. After, the resampling, all particle weights are then reset to

$$w_t^i = \frac{1}{N} \quad (21)$$

This enables the particle filter to estimate the robot's pose defiantly without growing a number of particles. However, resampling can delete good samples from the sample set, and in the worst case, the filter diverges. The decision on how to determine the point of time of the resampling is a fundamental issue. Liu introduced the so-called effective number of particles $N_{eff}$ to estimate how well the current particle set represents the true posterior. This quality is computed as

$$N_{eff} = \frac{1}{\sum_{i=1}^{N} w_k^i} \quad (22)$$

Where $w^i$ refers to the normalized weight of particle $i$ .The resampling process is operated whenever $N_{eff}$ is bellow a pre-defined threshold, $N_{tf}$ . Here $N_{tf}$ is usually a constant value as following

$$N_{tf} = \frac{3}{4} M \quad (23)$$

Where $M$ is number of particles.

## 5. A Modified Localization Based on Particle Filter

Particle Filter relies on importance sampling, i.e., it uses proposed distributions to approximate the posterior distribution. The most common choice of the proposed distribution that is used also in this paper is the probabilistic model of the states evolution, i.e., the transition prior $p(x_k^i | x_{k-1}^i)$ . Because the proposed distribution is suboptimal, there are two serious problems in particle filter. One problem is sample impoverishment, which occurs when the likelihood $p(z_k | x_k^i)$ is very narrow or likelihood lies in the tail of the proposed distribution $q(x_k^i | x_{k-1}^i, y_{1:k})$ . The prior distribution is effective when the observation accuracy is low. But it is not effective when prior distribution is a much broader distribution than the likelihood (such as Fig.4.). Hence, in the updating step, only a few particles will have significant importance weights.

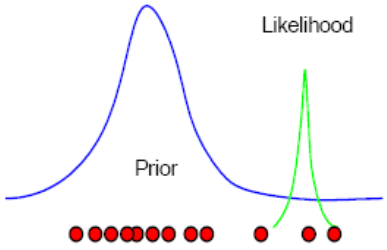


Fig.4 Prior and Likelihood

This problem implies that a large computational effort is devoted to update the particles with negligible weight. Thus, the sample set only contains few dissimilar particles and sometimes they will drop to a single sample after several iterations. As a result, important samples may be lost. Another problem of particle filter is the number of particles dependency that estimates the pose of the robot. If the number of particles is small, then there might not have been particles distributed around the true pose of the

robot. So after several iterations, it is very difficult for particles to converge to the true pose of the robot. For standard particle filter, there is one method to solve the problem. This is to augment the number of the particles. But this would make the computational complexity unacceptable. To solve these problems of particle filter, particle swarm optimization is considered to optimize the sampling process of the particle filter.

### 5.1 Particle Swarm Optimization

James Kennedy and Russell C.Eberhart [25] originally proposed the PSO algorithm for optimization. PSO is a population-based search algorithm based on the simulation of the social behavior of birds within a flock. PSO is initialized with a group of random particles and then computes the fitness of each one. Finally, it can find the best solution in the problem space via many iterations. In each iteration, each particle keeps track of its coordinates which are associated with the best solution it has achieved so far (pbest) and the coordinates which are associated with the best solution achieved by any particle in the neighboring of the particle (gbest). Supposing that the search space dimension is D and number particles is N, the position and velocity of the i-th particle are represented as $x_i = (x_{i1},...,x_{iD})$ and $v_i = (v_{i1},...,v_{iD})$ respectively.

Let $P_{bi} = [p_{i1},...p_{iD}]$ denote the best position which the particle $i$ has achieved so far, and $P_g$ the best of $P_{bi}$ for any $i = 1,...,N$ . The PSO algorithm could be performed by the following equations:

$$\vec{x}_i(t) = \vec{x}_i(t-1) + \vec{v}_i(t) \tag{24}$$

$$\vec{v}_i(t) = w\vec{v}_i(t-1) + c_1 r_1(\vec{P}_{bi} - \vec{x}_i(t-1)) + c_2 r_2(\vec{P}_g - \vec{x}_i(t-1)) \tag{25}$$

Where $t$ represents the iteration number and $c_1$, $c_2$ are the learning factors. Usually $c_1 = c_2 = 2$. $r_1, r_2$ are random numbers in the interval $(0,1)$. $w$ is the inertial factor, and the bigger the value of $w$ , the wider is the search range.

### 5.2 Localization based Multi Swarm particle filter

As discussed in the previous section, impoverishment occurs when the number of particles in the high likelihood area is low. We address this problem by intervening at Localization based on PF after the generation of the samples in prediction phase and before resampling. The aim is to move these samples towards the region of the state space where the likelihood is significant, without allowing them to go far away from the region of significant prior. For this purpose, we consider a multi objective function as follows:

$$F = F_1 + F_2 \tag{26}$$

The first objective consists of a function that is maximized at regions of high likelihood as follows:

$$F_1 = e^{-\frac{1}{2}(y_k - \hat{y}_k)^T [R]^{-1} (y_k - \hat{y}_k)} \tag{27}$$

Here, $R$ is the measurement noise covariance matrix, $\hat{y}$ is the predicted measurement and $y$ is the actual measurement. While the second objective, F2, is maximized at regions of high prior.

$$F_2 = e^{-\frac{1}{2}(x_k - \hat{x}_k)^T [Q]^{-1} (x_k - \hat{x}_k)} \tag{28}$$

where $Q$ is the measurement noise covariance matrix. We use an easy idea to solve this problem. The basic idea is that particles are encouraged to be at the region of high likelihood by incorporating the current observation without allowing them to go far away from the region of significant prior before the sampling process. This implies that a simple and effective method for this purpose is the using of PSO. In fact, by using PSO, we can move all the particles towards the region that maximizes the objective function $F$ before the sampling process. For this purpose, we consider a fitness function as follows:

$$\text{Fitness(k)} = \frac{1}{2}(y_k - \hat{y}_k) R^{-1} (y_k - \hat{y}_k)^{-1} + (x_k - \hat{x}_k^-) Q^{-1} (x_k - \hat{x}_k^-)^T \tag{29}$$

The particles should be moved such that the fitness function is optimal. This is done by tuning the position and velocity of the PSO algorithm. The standard PSO algorithm has some parameters that need to be specified before use. Most approaches use uniform probability distribution to generate random numbers. However it is difficult to obtain fine tuning of the solution and escape from the local minima using a uniform distribution. Hence, we use velocity updates based on the Gaussian distribution. In this situation, there is no more need to specify the parameter learning factors $c_1$ and $c_2$. Furthermore, using the Gaussian PSO the inertial factor $\omega$ was set to zero and an upper bound for the maximum velocity $v_{\max}$ is not necessary anymore [26]. So, the only parameter to be specified by the user is the number of particles. Initial values of particle filter are selected as the initial population of PSO. Initial velocities of PSO are set equal to zero. The PSO algorithm updates the velocity and position of each particle by following equations [26]:

$$\vec{x}_i(t) = \vec{x}_i(t-1) + \vec{v}_i(t) \quad (30)$$

$$\vec{v}_i(t) = |randn|(P_{pbest} - \vec{x}_i(t-1)) + |randn|(P_{gbest} - \vec{x}_i(t-1)) \quad (31)$$

PSO moves all particles towards particle with best fitness. When the best fitness value reaches a certain threshold, the optimized sampling process is stopped. With this set of particles the sampling process will be done on the basis of the proposed distribution. The Corresponding weights will be as follows:

$$w_k^i = w_{k-1}^i p(y_k \mid x_k^i) \quad (32)$$

Where

$$p(y_k \mid x_k^i) = \frac{1}{\sqrt{(2\pi)\,|R|}} \exp \{-\frac{1}{2}(y_k - \hat{y}_k)^T [R]^{-1} (y_k - \hat{y}_k)\} \quad (33)$$

Flowchart proposed algorithm is shown in Fig.5.

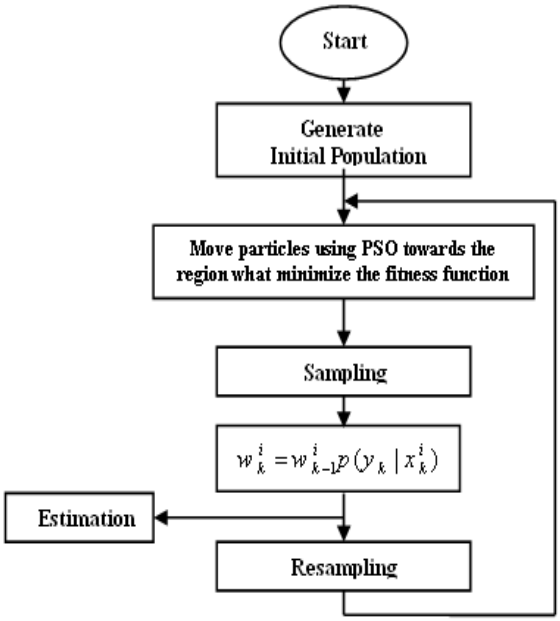



Fig.5 Modified particle algorithm

The pseudo code of Multi Swarm particle filter is as follows:

**Step1. General initial population initialization**

1) Initialize particle velocity
2) Initialize particle position
3) Initialize particle fitness value
4) Initialize pbest and gbest

**Step2. Move particles using PSO towards the region what minimize the fitness function**

Adjust the speed and location of particles

**Step3.** Sampling

**Step4.** Assign the particle a weight

$$w_k^i = w_{k-1}^i p(y_k \mid x_k^i)$$

**Step5.** The normalized weights

$$w_k^i = \frac{w_k^i}{\sum_{i=1}^{N} w_k^i}$$

**Step5.** Resampling

The resampling is operated whenever $N_{eff}$ is bellow a predefined threshold

**Step6.** Prediction

Each pose is passed through the system model

**Setp7.** Increase time $k$ and return to step 2.

## 6. Implementation and Results

Simulation experiments have been carried out to evaluate the performance of the proposed approach in comparison with the classical method. The proposed solution for the Localization problem has been tested for the benchmark environment, with varied number and position of the landmarks. Fig.6 shows the robot trajectory and landmark location (Map of environment). The star points (*) depict the location of the landmarks that are known and stationary in the environment.

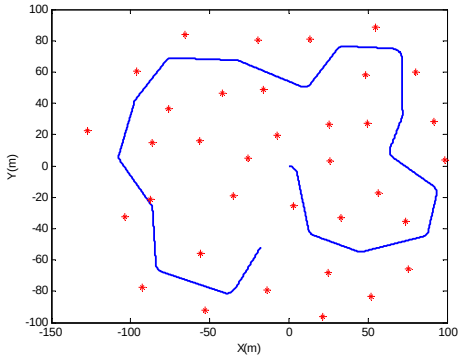



Fig.6 The experiment environment: The star point "*" denote the landmark positions (Map) and blue line is the path of robot.

The initial position of the robot is assumed to be $x_0 = 0$. The robot moves at a speed of 3m/s and with a maximum steering angle of 30 degrees. Also, the robot has 4 meters wheel base and is equipped with a range-bearing sensor with a maximum range of 20 meters and a 180 degrees frontal field-of-view. The control noise is $\sigma_v = 0.3$ m/s and $\sigma_\gamma = 3^o$. A control frequency is 40 HZ and observation scans are obtained at 5 HZ. The measurement noise is 0.2 m in range and $1^o$ in bearing. Data association is assumed known. The performance of the two algorithms can be compared by keeping the noises level (process noise and measurement noise) and varying the number of

particles. Fig.7 to Fig.12 shows the performance of the two algorithms. The results are obtained over 50 Monte Carlo runs. As observed, localization based on multi swarm particle filter (PFPSO) is more accurate than the localization based on PF. Also, performance of the proposed method does not depend on the number of particles while the performance of localization based on PF highly depends on the number of particles. For very low numbers of particles, localization based on PF diverges while the proposed method is completely robust. This is because PSO in the proposed method places the particles in the high likelihood region. In addition, we observed that the proposed method requires fewer particles than localization based on PF in order to achieve a given level of accuracy for state estimates.

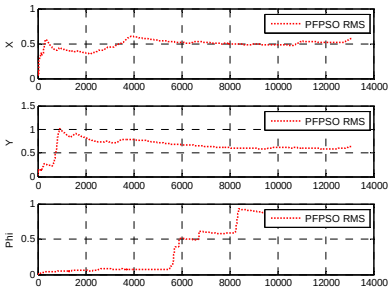



Fig.7 RMS error of localization based on PFPSO and number of particles is 5

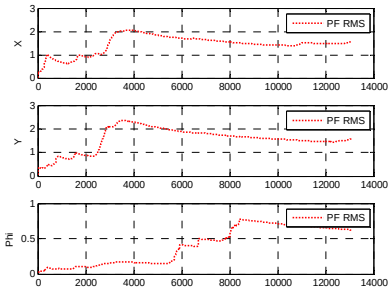



Fig.8 RMS error of localization based on PF and number of particles is 5

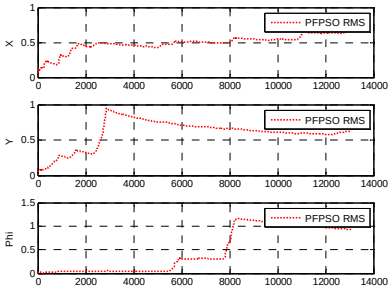



Fig.9 RMS error of localization based on PFPSO and number of particles is 10

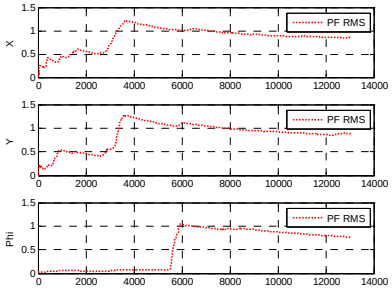



Fig.10 RMS error of localization based on PF and number of particles is 10

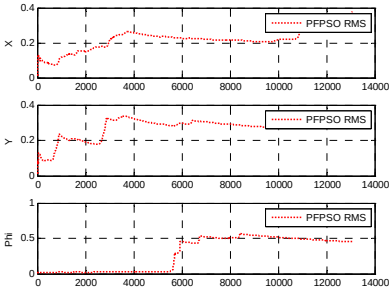



Fig.11 RMS error of localization based on PFPSO and number of particles is 20

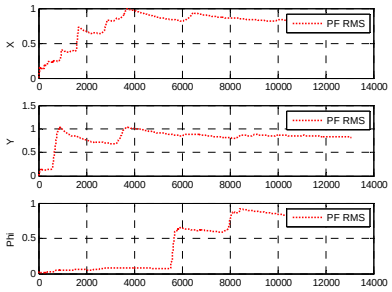



Fig.12 RMS error of localization based on PF and number of particles is 20

## Conclusion

This paper proposed a new method for the accurate localization of a mobile robot. The approach is based on the use of PSO for improving the performance of the particle filter. The problem of localization based on PF is that it degenerates over time due to the loss of particle diversity. One of the main reasons for loosing particle diversity is sample impoverishment. It occurs when likelihood lies in the tail of the proposed distribution. In this case, most of participles weights are insignificant. This paper presents a modified localization based on PF by soft computing. In the proposed method, a particle filter based on particle swarm optimization is presented to overcome the impoverishment of localization based on particle filter. Finally, Experimental results confirm the

effectiveness of the proposed algorithm. The main advantage of our proposed method is its more consistency than the classical method. This is because in our proposed method, when motion model is noisier than measurement, the performance of the proposed method outperforms the standard method. The simulation results show that state estimates from the multi swarm particle filter are more accurate than the Particle filter.

## References

[1] F.Kong, Y.Chen, J.Xie, Gang, "Mobile robot localization based on extended kalman filter", Proceedings of the 6th World Congress on Intelligent Control and Automation, June 21-23, 2006.

[2] J.Kim, Y.Kim, and S.Kim, "An accurate localization for mobile robot using extended kalman filter and sensor fusion", Proceeding of the 2008 International Joint Conference on Neural Networks, 2008.

[3] Tran Huu Cong, Young Joong Kim and Myo-Taeg Lim, "Hybrid Extended Kalman Filter-based Localization with a Highly Accurate Odometry Model of a Mobile Robot", International Conference on Control, Automation and Systems, 2008.

[4] Sangjoo Kwon, Kwang Woong Yang and Sangdeok Park, "An Effective Kalman filter Localization Method for mobile Robots", Proceeding of the IEEE/RSJ, International Conference on Intelligent Robots and Systems, 2006.

[5] W.Jin, X.Zhan, "A modified kalman filtering via fuzzy logic system for ARVs Localization", Proceeding of the IEEE, International Conference on Mechatronics and Automation, 2007.

[6] G.Reina, A.Vargas, KNagatani and K.Yoshida "Adaptive Kalman Filtering for GPS-based Mobile Robot Localization", in Proceedings of the IEEE, International Workshop on Safety, Security and Rescue Robotics, 2007.

[7]Y.Xia , Y.Yang," Mobile Robot Localization Method Based on Adaptive Particle Filter", C. Xiong et al. (Eds.): ICIRA 2008, Part I, LNAI 5314, pp. 963–972, Springer-Verlag Berlin Heidelberg, 2008.

[8]J.Zheng-Wei,G.Yuan-Tao, "Novel Adaptive Particle Filters In Robot Localization", Journal of Acta Automatica Sinica,Vol.31,No.6,2005.

[9]S.Thrun," Particle Filters in Robotics", In Proceedings of Uncertainty in AI (UAI), 2002.

[10]Jeong Woo, Young-Joong Kim, Jeong-on Lee and Myo-Taeg Lim," Localization of Mobile Robot using Particle Filter", ICE-ICASE International Joint Conference ,2006.

[11]D. Zhuo-hua, F. Ming,, C. Zi-xing,, YU Jin-xia, " An adaptive particle filter for mobile robot fault diagnosis", Journal of Journal of Central South University, 2006.

[12]F. Chausse, S.Baek, S.Bonnet, R.Chapuis, J.Derutin," Experimental comparison of EKF and Constraint Manifold Particle Filter for robot localization", Proceedings of IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems Seoul, Korea, August, 2008.

[13]J.Woo, Y.Kim, J.Lee ,M.Lim, " Localization of Mobile Robot using Particle Filter", SICE-ICASE International Joint Conference , 2006.

[14]G.Cen, N.Matsuhira, J.Hirokawa, H.Ogawa, I.Hagiwara, " Mobile Robot Global Localization Using Particle Filters", International Conference on Control, Automation and Systems, 2008.

[15]S.Thrun, D.Fox, W.Burgard, F.Dellaert," Robust Monte Carlo localization for mobile robots", Journal of Artificial Intelligence, 2001.

[16]Thrun, S., Fox, D., Burgard, W., Dellaert, F., " Robust monte carlo localization for mobile robots", Artificial Intelligence, 2001.

[17]D. Fox, "Adapting the sample size in particle filters through KLD-sampling", The International Journal of Robotics Research, 2003.

[18] D. Fox., "KLD-sampling: Adaptive particle filters and mobile robot localization", In Advances in Neural Information Processing Systems , 2001.

[19] D.Simon, " Optimal State Estimation Kalman, $H_\infty$ and Nonlinear Approaches ", John Wiley and Sons, Inc, 2006

[20] M.Sanjeev Arulampalam, S.Maskell, N.Gordon, and Tim Clapp, "A Tutorial on Particle Filters for Online Nonlinear/Non-Gaussian Bayesian Tracking", IEEE Transactions on Signal Processing (S1053-587X) 50(2), 174–188, 2002.

[21] Liang Xiaolong,Feng Jinfu and Li Qian Lu Taorong, Li Bingjie, " A Swarm Intelligence Optimization for Particle Filter" ,Proceedings of the 7th World Congress on Intelligent Control and Automation June 25 - 27, Chongqing, China, 2008.

[22] Guofeng Tong, Zheng Fang, Xinhe Xu," A Particle Swarm Optimized Particle Filter for Nonlinear System State Estimation", IEEE Congress on Evolutionary Computation Sheraton Vancouver Wall Centre Hotel, Vancouver, BC, Canada ,July 16-21, 2006.

[23] Jian Zhou, Fujun Pei, Lifang Zheng and Pingyuan Cui," Nonlinear State Estimating Using Adaptive Particle Filter", Proceedings of the 7th World Congress on Intelligent Control and Automation June 25 - 27, 2008.

[24]Gongyuan Zhang, Yongmei Cheng, Feng Yang, Quan Pan, " Particle Filter Based on PSO", International Conference on Intelligent Computation Technology and Automation, 2008.

[25] R.C. Eberhart, J. Kennedy, "A new optimizer using particle swarm theory", in: Proceedings of the Sixth International Symposium on Micromachine and Human Science, Nagoya, Japan, pp. 39–43, 1995.

[26] R. A. Krohling," Gaussian swarm: a novel particle swarm optimization algorithm", In Proceedings of the IEEE Conference on Cybernetics and Intelligent Systems (CIS), Singapore, pp.372-376, 2004.

**Ramazan Havangi** received the M.S. degree in Electrical Engineering from K.N.T.U University, Tehran, Iran, in 2004; He is currently working toward the Ph.D. degree in K.N.T.U University. His current research interests include Inertial Navigation, Integrated Navigation, Estimation and Filtering, Evolutionary Filtering, Simultaneous Localization and Mapping, Fuzzy, Neural Network, and Soft Computing.

**Mohammad Ali Nekoui** is assistant professor at Department of Control, Faculty of Electrical Engineering K.N.T.U University. His current research interests include Optimal Control Theory, Convex Optimization, Estimation and Filtering, Evolutionary Filtering, Simultaneous Localization and Mapping.

**Mohammad Teshnehlab** is professor at Department of Control, Faculty of Electrical Engineering, K.N.T.U University. His current research interests include Fuzzy, Neural Network, Soft Computing, Evolutionary Filtering, and Simultaneous Localization and Mapping.

IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 3, No 2, May 2010
ISSN (Online): 1694-0784
ISSN (Print): 1694-0814


# Development of Receiver Stimulator for Auditory Prosthesis

**K. Raja Kumar , P. Seetha Ramaiah**

**Dept of Computer Science and Systems Engg, Andhra University**
**Visakhapatnam, Andhra Pradesh, 530003,INDIA**

**Abstract**

***The Auditory Prosthesis (AP) is an electronic device that can provide hearing sensations to people who are profoundly deaf by stimulating the auditory nerve via an array of electrodes with an electric current allowing them to understand the speech. The AP system consists of two hardware functional units such as Body Worn Speech Processor (BWSP) and Receiver Stimulator. The prototype model of Receiver Stimulator for Auditory Prosthesis (RSAP) consists of Speech Data Decoder, DAC, ADC, constant current generator, electrode selection logic, switch matrix and simulated electrode resistance array. The laboratory model of speech processor is designed to implement the Continuous Interleaved Sampling (CIS) speech processing algorithm which generates the information required for electrode stimulation based on the speech / audio data. Speech Data Decoder receives the encoded speech data via an inductive RF transcutaneous link from speech processor. Twelve channels of auditory Prosthesis with selectable eight electrodes for stimulation of simulated electrode resistance array are used for testing. The RSAP is validated by using the test data generated by the laboratory prototype of speech processor. The experimental results are obtained from specific speech/sound tests using a high-speed data acquisition system and found satisfactory.***

***Keywords:*** *Receiver-stimulator, auditory prosthesis, microcontroller, Transcutaneous RF link.*

## 1. Introduction

The cochlear implant or Auditory Prosthesis (AP) has recently emerged as clinically acceptable prosthesis for aiding people suffering from a profound to total sensorineural hearing loss [1]. Several types of electronic hearing prostheses are now in widespread use by large number of people around the world. Today high performance and highly reliable multi channel auditory Prosthesis are available from various vendors such as Nucleus, Clarion, and Med-El etc [1]. These devices are expensive and not affordable by developing countries like India, China, Pakistan etc. [2-3]. Our aim is to design and develop low cost, high performance and highly reliable Auditory Prosthesis for use in developing countries. The developed system comprises the following main components: Body-Worn Speech Processor, a transcutaneous RF link, the prototype model of a 12 channel Receiver Stimulator for Auditory Prosthesis, and an array of electrodes.

The laboratory model of BWSP is a programmable speech processor [4-5] that implements 4 to 8 channels CIS algorithm [6-7] based on the number of active electrodes. The selection of active electrodes of each patient is determined by using the Clinical Programming Software (CPS) [8]. It can implement 4/5/6/7/8 CIS algorithm in which the frequency band of 200-6600Hz is logarithmically distributed to 4/5/6/7/8 bands/channels and set the patient specific compression values such as threshold and most comfort levels. It encodes the processed speech/sound information protocol specified in BWSP. The encoded speech data with Amplitude Shift Key (ASK) modulation is transmitted via radio-frequency (RF) transmitter, which is coupled to the Auditory Prosthesis via a transcutaneous inductive link. The ASK demodulator [9-11] in the prototype RSAP demodulates the incoming RF signal and reconstructs the encoded speech data. The encoded speech data is decoded by Speech Data Decoder and delivers electric current pulses with specified parameters to the selected channel in the electrode array.

The design as well as development of the prototype model of a 12 channel Receiver stimulator for Auditory Prosthesis is addressed in the present paper, covering both the Hardware and Software parts. The Hardware part covers the design and implementation of Speech data Decoder and Microcontroller interfacing between various analog and digital circuitry. The software part covers the embedded programs written for implementation of Speech data stimulation and impedance telemetry that sends back to speech processor the encoded electrode impedance data of each channel via RF transcutaneous inductive link.. It also covers the protocol implementation for Speech data stimulation from the data generated by the speech encoder in BWSP and the

protocol used in the impedance telemetry for identifying the patient specific active electrode contacts.

features of DS89c420 are used for real-time speech processing by the BWSP. Speech Decoder performs two essential functions such as Speech Processing and Impedance Telemetry. At any instant of time, Speech Data

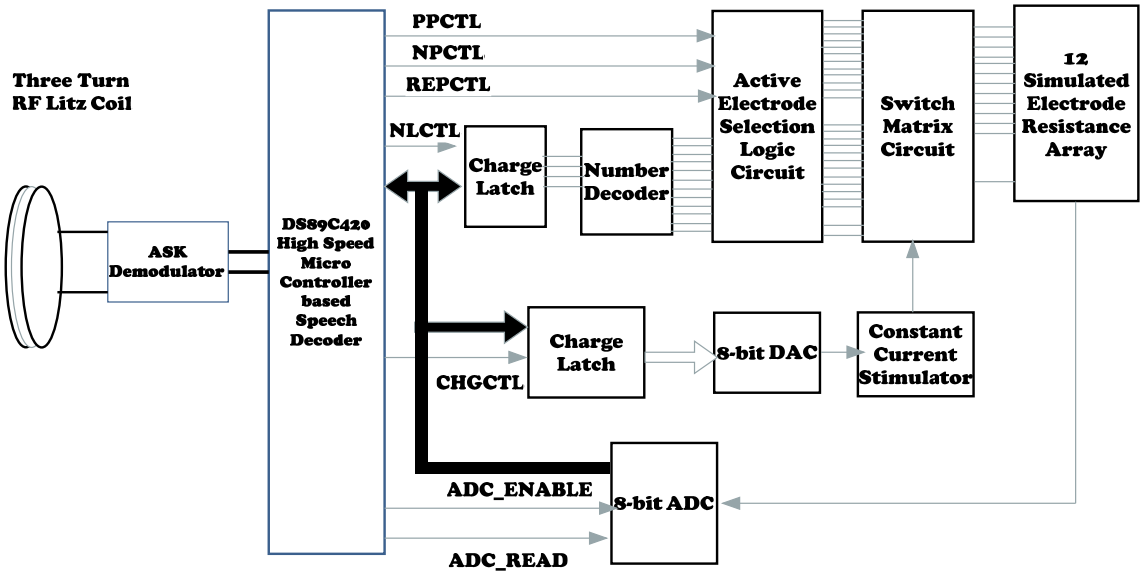



Figure-1: Block diagram of the Receiver Stimulator for Auditory Prosthesis.

## 2. Hardware Description

The Laboratory model for RSAP that consists of RF ASK receiver, Stimulus Generator, active electrode decoder, Stimulus buffer, 8-bit Digital to Analog Converter, constant current generator, active electrode selection logic driver, switch matrix and 12 simulated electrode resistance arrays is shown in figure-1.

The ASK Demodulator receives the high-speed serial data at around 172 Kbps rate from the 2 turn secondary Platinum iridium (Pt-Ir) coil and reconstructs corresponding digital signal of encoded speech data and fed to the serial receive input of Speech Decoder. Demodulation of ASK signal is a three stage process which consists of envelope detector, threshold detector and load driver. Envelope detector is simple diode detector which extracts low-high signal variations of 4MHz Carrier wave with a dc offset. The threshold detector is Schmitt trigger circuit which detects whether it is logic '0' or logic '1'. Final stage of Load driver is a buffer consists of two cascaded inverters to get noise free digital TTL signal.

Speech Data Decoder based on Dallas Semiconductor's 8-bit ultra-high-speed flash microcontroller DS89C420 that executes one instruction per clock cycle with 33MHz clock, meeting the demand of real time processing of the speech signals. The High-speed and High performance Decoder will perform any one of these two functions. Speech decoder receives the data bytes as per the protocol adapted for BWSP. The I/O pins of DS89C420 are configured as follows: one port for bi-directional port, one output port for required control signals.

The port P1 of DS89C420 is connected to two 8-bit edge-triggered D-type flip-flops as buffers with the output-control (OC) input. These two buffers are controlled by the two control signals: Number Latch Control (NLCTL) and Charge Control (CHGCTL) which are generated by the DS89C420 microcontroller for enabling or transferring the information to the respective buffer whenever it is needed.

Only 4-bit output of number latch is connected to 4-to-16 line decoder with latched inputs. Only 12 out of 16 output lines are used for selection of electrode lines. The 12 line outputs of the Number decoder are connected to the Active Electrode Selection Logic Circuit, which is a combinational logic circuit used to generate two pairs of control signals for 12 switches, one pair for the positive side of the bi-phasic pulse generation and other pair for negative side of the pulse generation. In addition to this, it generates switch control signals for reference electrode.

The 8-bit data stored in the charge buffer specifies amount of stimulus current amplitude to be stimulated. The output of the charge buffer is connected to 8-bit Digital to Analog Converter DAC0800. The DAC converts the 8-bit digital value into corresponding analog voltage as output. The

output of the DAC is connected to the Constant Current Stimulator which gives the constant current according to the input voltage. It maintains the constant current for load resistance up to 10K ohms. Constant current generator generates the current for 0 to 1mA for the input of 8-bit data in the charge buffer ranging from 00 to FF hexadecimal value. This constant current is given as stimulus to the switch matrix driver. Now the current flows across the electrode contact with reference to the reference electrode based on the selected electrode. By properly sending the required control signals to the switches foe closing, the constant current flows across the selected electrode contacts with respect to the reference electrode.

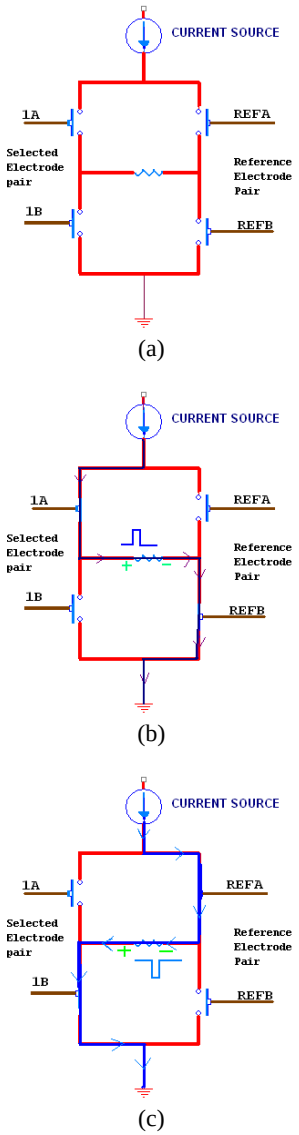



Fig.2 (a) Generation of biphasic pulses using H-matrix architecture. (b) Positive amplitude generation and (c) negative amplitude generation.

The one stage of electrode switching matrix arrangement is as shown in figure-2a based on H-architecture [12]. The H-architecture switches between two opposite pairs of analog switches delivering stimulation current in one direction by closing a pair of switches of the selected electrode from one group (e.g. 1A) and the reference electrode (REFB) and other direction delivering another pair of the selected electrode from second group (e.g. 1B) and the reference electrode (REFA). This circuit switching allows the delivery of the stimulus charge based on the current flow through the simulated electrode resistances. Figure -2b shows the generation of positive amplitude of bi-phsasic pulse as the current flows from the path 1A – resistor-REFB path and Figure-2C shows the negative amplitude of bi-phasic pulse as the current passes though the path REFA-resistor-1B.

## 3. Software Description

The software design is based on top-down approach that identifies the major components of the system, decomposing them into their lower-level components and iterating until the desired level of details is achieved. The two important functional modules such as Speech Processing module and Impedance Telemetry module are described by using the flowchart as shown in figure-3.

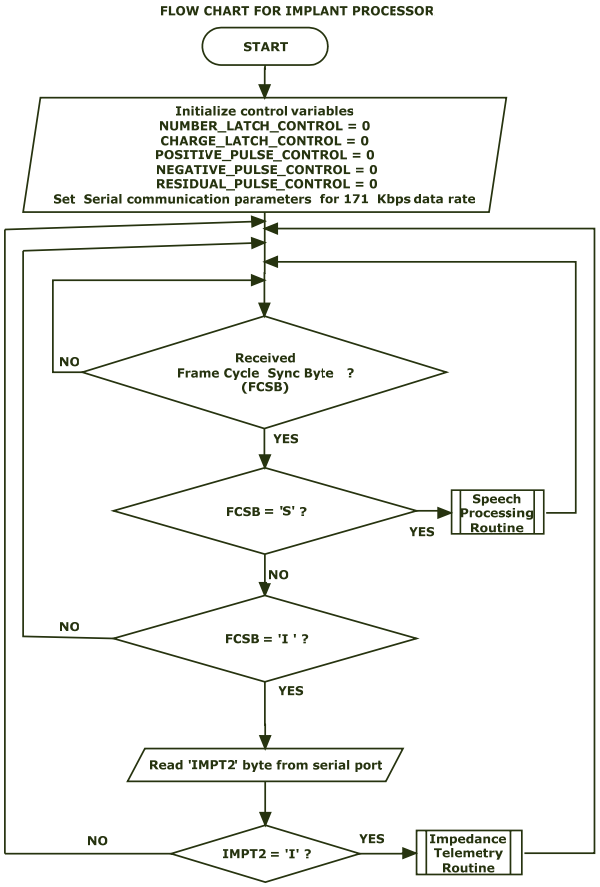



**Figure-3: Flowchart for the overall function of RSAP**

The main function of Speech Processing module is to receive stimulus information of one electrode at a time that contains electrode number and electrode charge and stimulate the corresponding electrode with given charge in the stimulus information. The function of Impedance Telemetry module function is to read the impedance of the electrodes and send this information to Impedance Telemetry module. This paper covers the speech data processing function of speech data decoding and stimulation.

The embedded software program in the DS89C420 microcontroller performs the tasks such as Speech decoder and Impedance telemetry processing. The NLCTL and CHGCTL are initialized to Logic '1', Pulse control signals are initialized to logic '0' and the parameters for the serial port of the microcontroller are initialized to 172Kbps baud rate. The reconstructed serial data from the ASK demodulator is received by the receive pin (rx) of DS89C420. The data format received from the BWSP is shown figure-4.

| FCSB | EL1 | CH1 | EL2 | CH2 | …. | EL8 | CH8 |
|---|---|---|---|---|---|---|---|

Figure-4: Protocol format sent by the BWSP

The data format contains Frame Cycle Sync Byte (FCSB), active eight pairs of electrode numbers ($EL_i$) and electrode charge ($CH_i$) for 8 electrodes. The BWSP sends the speech data in data format in figure-4 continuously. The RSAP takes two bytes of information that contains the electrode number and the charge. It places the first byte first on the port and issues the NLCTL signal enabling the buffer to load the new value which is used as input to the electrode selection circuit and places the next byte into the port that contains the charge to be stimulated and issues a pulse to the CHGCTL signal which loads the new values served as input to the DAC and current stimulator. By setting the timer values of 14us for both positive and negative pulses of biphasic pulse with associated control signals, biphasic pulses are generated .The operation of overall system is represented by a flowchart as shown in figure-5.

## 3.1. Impedance Telemetry functions:

The impedance telemetry function is explained with the help of flowchart as shown in Figure-6. If the received FCSB is 'I' it indicates that the function is impedance telemetry function. It reads the impedance of each channel by stimulating each channel with "0xFF", using ADC and store each resistance/impedance value in the internal memory. After reading resistance ($RES_i$) of each channel, it prepares the response byte as shown in figure-7.

| RES | EL1 | RES1 | …. | EL8 | RES8 |
|---|---|---|---|---|---|

Figure-7: Response format sent by the IRS

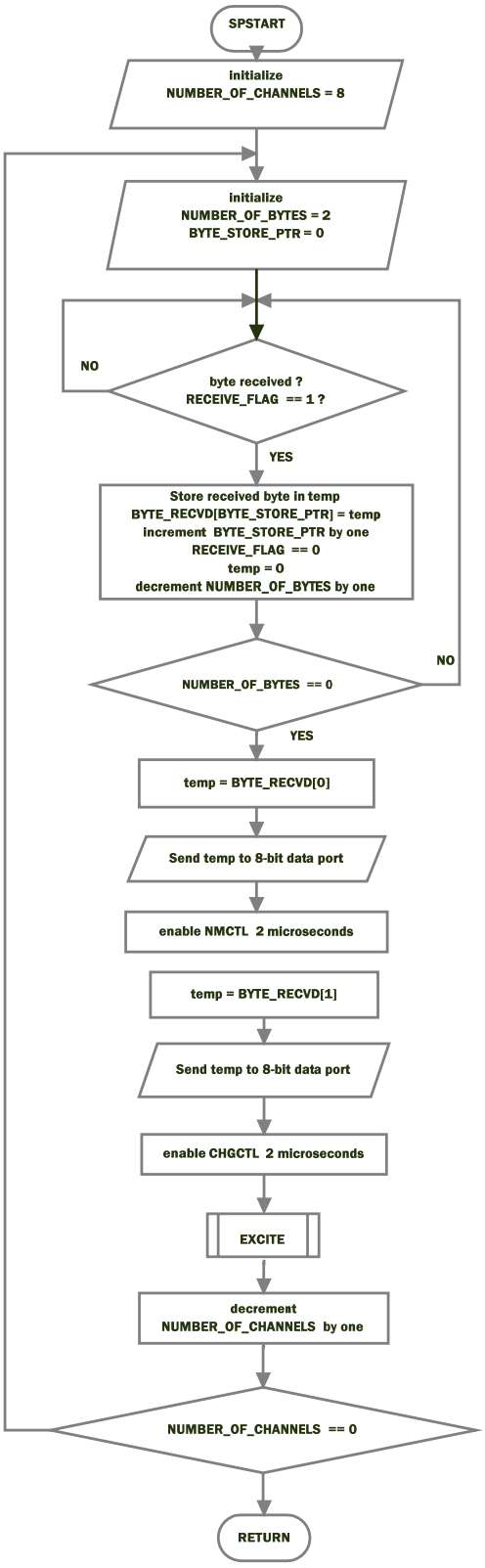



**Figure-5: Flowchart for the Speech Processing routine**

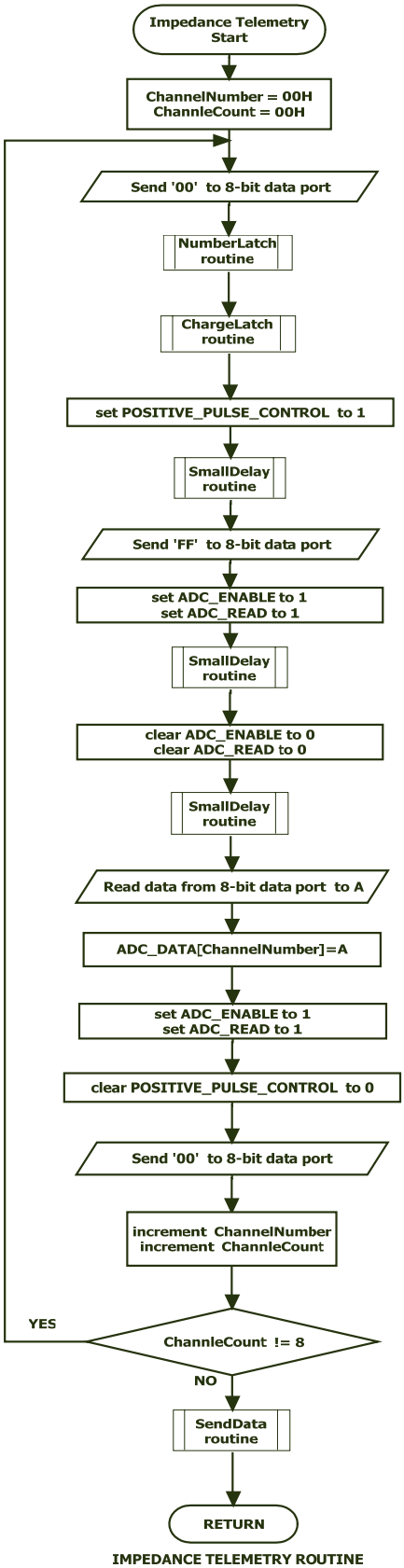



**Figure-6: Flowchart for impedance telemetry routine**

## 4. Results

The functionality of the prototype model is tested and validated using test data samples and real-time speech samples. Test samples for speech processing routine are generated by using laboratory model of BWSP. The BWSP is initially programmed for 8 channel Speech processor and observed satisfactory performance. The several identified test data patterns are applied to the RSAP and observed the respective electrode stimulation signals at simulated electrode resistance array. One of the sample test pattern is by sending equal charge to all the electrodes. The corresponding stimulation outputs across the electrodes with different load resistance values are shown in figure-8.

Impedance telemetry (IMT) function is tested by using laboratory model of IMT module after issuing the impedance telemetry command to RSAP and read the resistance values sent by the RSAP as per the response format, stored in the internal memory of IMT. By using UP/DOWN key of the IMT module, the resistance values of all 12 electrodes are observed. The process is repeated several times by varying the simulated resistances of electrodes and observed the resistance values as expected. This RSAP is also tested with the CPS software and IMT module. By using the CPS software, the resistance values of each patient are read and stored in the database for future use. CPS software is also used to generate the stimulus pulses of varying amplitudes to determine the patient threshold and most comfort values.

## 5. conclusions

The Microcontroller Based Prototype model of RSAP for Hearing Impaired research has been developed. The system has been tested by using simulated test data. The Laboratory model of Body Speech Processor is used to test 8-channel auditory prosthesis by sending the encoded ASK modulated serial data bits to RSAP for real-time speech signal. The RSAP is validated by stimulating the selected electrodes as the simulated electrode resistance array and observed satisfactory results. The impedance telemetry feature is also tested using the Laboratory model of IMT module. Individual channels are stimulated with associated charge to determine the patient thresholds. The conversion from prototype model to CMOS ASIC model is under progress.

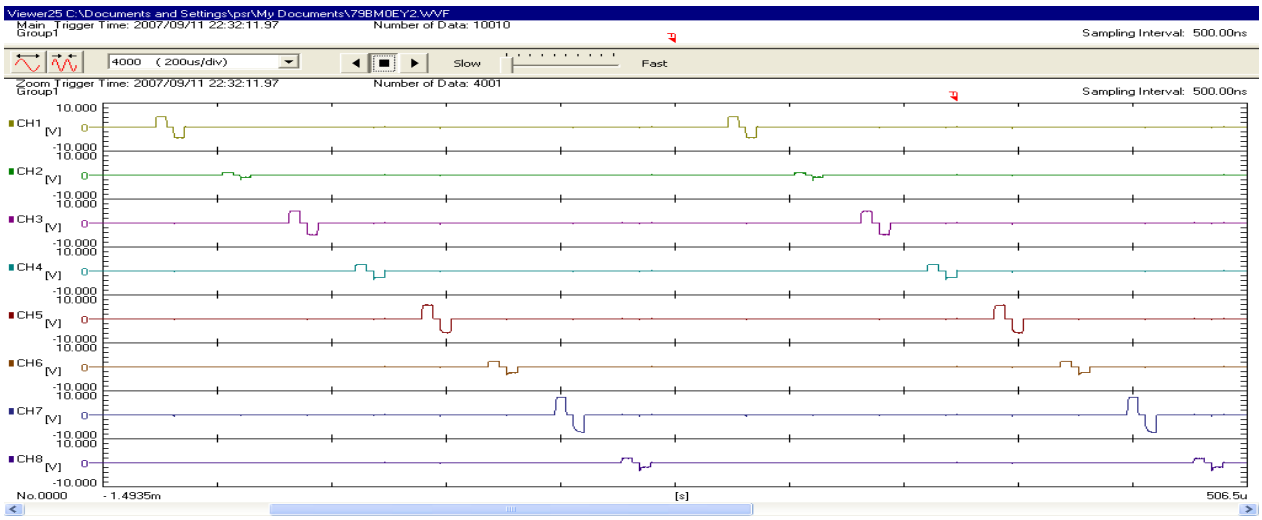



**Figure-8:** Responses of the 8 electrodes across the simulated electrode resistance array

**Acknowledgments**

The authors wish to thank Dr. V. Bhujanga Rao, Director, NSTL, Visakhapatnam who provided suitable hardware and software tools for successful completion of laboratory prototype of Receiver Stimulator for Auditory Prosthesis. This work is supported by defense R & D organization – NSTL-Visakhapatnam, INDIA under the contract no. NSTL/WI/CI dated 26 February 2009.

## References

[1] Blake S. Wilson, "The Surprising Performance of Present-Day Cochlear Implants", IEEE Tran Biomedical Engg, vol. 54, no. 6, pp . 969 -972, June 2007

[2] Soon Kwan An, Se-Ik Park, Sang Beom Jun, Choong Jae Lee, Kyung Min Byun, Jung Hyun Sung, Blake S. Wilson, "Design for a Simplified Cochlear Implant System", IEEE Tran Biomedical Engg, vol. 54, no. 6, pp . 973-982,June 2007

[3] B. S. Wilson, S. Rebscher, F. G. Zeng, R. V. Shannon, G. E. Loeb,D. T. Lawson, and M. Zerbi, "Design for an inexpensive but effective cochlear implant," Otolaryngol.—Head Neck Surg., vol. 118, no. 2, pp. 235–241, Feb. 1998.

[4] K. Raja Kumar and P. Seetha Ramaiah, "DSP and Microcontroller based Speech Processor for Auditory Prosthesis", Proceedings of the 14th International Conference on Advanced Computing and Communication, ADCOM-2006, December 20-23, 2006, pp.518-522.

[5] K. Raja Kumar and P. Seetha Ramaiah, "Programmable Digital Speech Processor for Auditory Prostheses", Proceedings of IEEE TENCON 2008 on Innovative Technologies for Societal Transformation, TENCON-2008, Nov.18-21, 2008.

[6] B. S.Wilson, C. C. Finley, D. T. Lawson, R. D.Wolford, and M. Zerbi, "Design and evaluation of a continuous interleaved sampling (CIS) processing strategy for multichannel cochlear implants," J Rehabil. Res. Dev., vol. 30, no. 1, pp. 110–116, 1993.

[7] B. Wilson, C. Finley, D. Lawson, R. Wolford, D. Eddington, and W. Rabinowitz, "Better speech recognition with cochlear implants," Nature, vol. 352, pp. 236-238, July 1991.

[8] K. Raja Kumar and P. Seetha Ramaiah, "Personal Computer Based Clinical Programming Software for Auditory Prostheses", Journal of Computer Science, Vol. 5 Issue.8, pp. 589-595, 2009

[9] G. A. Kendir, W. Liu, G. Wang, M. Sivaprakasam, R. Bashirullah, M. S. Humayun, and J. D. Weiland, "An optimal design methodology for inductive power link with class-E amplifier," IEEE Trans. Circuits Syst.I, Reg. Papers, vol. 52, no. 5, pp. 857–866, May 2005.

[10] Clemens M. Zierhofer, Ingeborg J. Hochmair-Desoyer, and Erwin *S.* Hochmair , "Electronic Design of a Cochlear Implant for Multichannel High-Rate Pulsatile Stimulation Strategies", IEEE Trans Rehab. Eng, vol. 3, no. 1, pp 112-116 ,March *1995*

[11] K. Raja Kumar and P. Seetha Ramaiah, "Microcontroller based receiver stimulator for auditory prosthesis", Proceedings of IEEE TENCON 2008 on Innovative Technologies for Societal Transformation, TENCON-2008, Nov.18-21, 2008.

[12] Chih-Kuo Liang, Gin-Shu Young, Jia-Jin Jason Chen and Chung-Kai Chen, "Microcontroller-based implantable Neuromuscular stimulation system with Wireless power and data transmission for Animal experiments", Journal of the Chinese Institute of Engineers, Vol. 26, No. 4, pp. 493-501,2003.

**K. Raja Kumar** received his B.E Degree in Electronics and Communication Engineering in 1998 and M.Tech Degree Computer Science and Technology in 2000 from Andhra University, Visakhapatnam-India. He is presently working as an Assistant Professor and pursuing for Ph.D. in the area of computerized bionic implants in the Department of Computer Science and Systems Engineering, A.U. College of Engineering, Andhra University. He has published five International Conference papers. His areas of research include Embedded Systems, Signal Processing algorithms, Real-time systems and VLSI design. He is member of IEEE Computer society.

**Dr. P. Seetha Ramaiah** received his PhD in Computer Science from Andhra University in 1990. He is presently working as a Professor in the department of Computer Science and Systems Engineering, A.U. College of Engineering, Visakhapatnam, INDIA. He is the Principal Investigator for several Defence R&D projects and Department of Science and Technology projects of the Government of India in the areas of Embedded Systems and robotics. He has published seven journal papers, and presented Fifteen International Conference papers in addition to twenty one papers at National Conferences in India. His areas of research include Safety-Critical Computing- Software Safety, Computer Networks, VLSI and Embedded Systems, Real-Time Systems, Microprocessor-based System Design, Robot Hand-Eye Coordination, Signal Processing algorithms on fixed-point DSP processors, Bio-Electronics Systems.

IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 3, No 2, May 2010
ISSN (Online): 1694-0784
ISSN (Print): 1694-0814


# An Efficient Ball Detection Framework for Cricket

**B.L. Velammal[1] and P. Anandha Kumar[2]**

**[1]Department of CSE, Anna University,
Chennai, Tamilnadu 600025, India**

**[2]Department of IT, MIT Campus, Anna University,
Chennai, Tamilnadu 600025, India**

**Abstract**

Ball Detection and Tracking in Cricket image sequences has become a growing and challenging issue, with the rising popularity of Sports analysis. To identify the ball in cricket is very important for event recognition. It is also useful for summarization. Lot of methods has been proposed for ball detection in Soccer videos but ball detection in cricket is more challenging than Soccer because of the smaller ball and the ball deforms while moving. An anti-model approach is used to eliminate non-ball objects and remaining objects are identified as ball-objects. Region Growing segmentation is chosen for segmentation. After segmentation, the ball and non-ball objects are classified using the shape properties. The non-ball objects are eliminated and the resulting frames consists of only ball objects or ball-candidates. The ball candidates are to be processed further to detect the ball. This method eliminates false alarms in ball detection.

***Keywords*****:** *Segmentation, Ball Detection, anti-model approach, and ball-candidates.*

## 1. Introduction

Ball detection in cricket has achieved very much significance with the introduction of twenty - twenty cricket matches. Automatic ball recognition in the television image sequences is a fundamental task to be solved. Ball Detection in Cricket domain is very challenging as a great number of problems have to be managed such as occlusions, shadows, objects similar to the ball, etc. The Ball which is hit can be distorted in shape as it moves with certain velocity or it can be occluded with the pitch or outfield, the movement of the camera and the size of the ball is relatively small when compared to all other objects in a frame. Because of these it is difficult to detect the ball and find its position in the frame [4].

Instead of going for conventional ball detection methods, a different approach is adopted. To search and find the ball in an entire frame is a tedious task. First the frame is segmented and then the ball-candidates alone are generated from the frame. This is done by classification of the frame. The non-ball candidates are removed and only ball-candidates is used for further processing. Informally, the key idea behind this strategy is while it is very challenging to achieve high accuracy in locating the precise location of the ball, it is relatively easy to achieve very high accuracy in locating the ball among a set of ball-like candidates. An important feature of this paper is that it can be extended for detecting events. Interesting events such as boundaries and sixes hit etc. can be found out by extending this idea. Previously ball detection has been carried out for other sports like Tennis, Basket Ball and Soccer. The idea can also be further extended to obtain automatic highlights extraction in sports.

In the last decade object detection and tracking become very popular because of its applicability to daily problems and ease of production, e.g. surveillance cameras, adaptive traffic lights with object tracking, plane detection. The superiority of object-tracking to object recognition became apparent after the development in the video processing and motion estimation. Although object detection and tracking using motion vectors is a very powerful method, it fails to give a robust and reliable answer all the time.

Object detection is necessary for surveillance applications, for efficient video compression, for smart tracking of moving objects, for automatic target recognition (ATR) systems and for many other applications. The convergence of computer vision and multimedia technologies has led to opportunities to develop applications for automatic sports video analysis, including content based indexing, retrieval and visualization.

With the advent of interactive broadcasting and interactive video reviewing [14], automatic sports video indexing would allow sports fans to access a game in the way that they like rather than watch a game in a sequential manner.

Existing methods that do direct ball detection in Cricket are limited to several inherent difficulties associated like:

ISSN (Online): 1694-0784
ISSN (Print): 1694-0814

- the very small size of the ball when compared to other sports like Soccer
- the ball is not exactly similar in all the frames due to high velocity it attains when the batsmen hits the ball and the bowlers bowls the ball at speeds as high as 160 kmph, and it is difficult to find the exact shape of the ball
- the presence of many ball-like objects in a frame and occlusion of the ball (say, by a player) in frames[1]

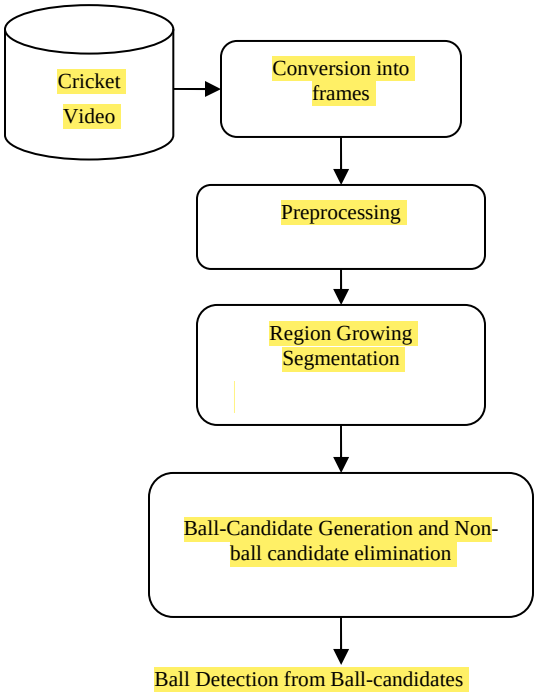



Fig. 1. Segmentation and ball detection in Cricket image sequences

## 2. Ball Detection Process

In the ball detection algorithm, the first step is to preprocess the video to remove noise, to enhance contrast etc. Before that, the video should be converted into frames for processing. Each and every frame is processed and the video is reconstructed from the frames in the final step. After preprocessing, segmentation is performed using the seeded Region Growing Algorithm. Various sieves such as shape, size, and color are used to sieve out the non-ball objects and the remaining objects are referred as ball-candidates which satisfy all the properties defined by the sieves.

### 2.1. Conversion of video to Frames

Instead of reading the video as it is, we converted the video into frames. Reading the video and directly processing the video is a tedious task and requires lot of memory and can work only on systems with high configurations. In a video clip each and every frame should be grabbed at a fixed frame capture rate. Now, each and every frame is an individual image and we can apply all image processing algorithms to these captured frames which is a major advantage and the size of the video does not matter here and no need of specific memory requirements. After performing all image processing operations, the video can be reconstructed from the frames by simple looping operations.

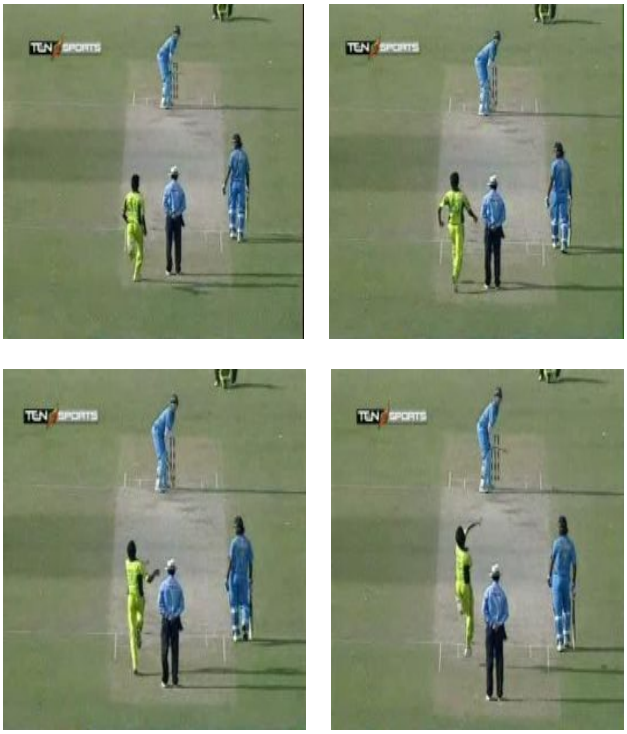


Fig. 2. Frame Sequences captured

### 2.2. Preprocessing of the Cricket video

Given any video frame, we need to preprocess the video to remove noise. With median filtering the value of the output pixel is determined by the median of the neighborhood pixels. The median is less sensitive than the mean to extreme values. Median filtering is therefore a good choice to remove these extreme values without reducing the sharpness of the image. Median filtering is applied since the goal is to remove the noise and preserve the edges, as it may carry useful information. The contrast should also be enhanced as a part of preprocessing. Proper preprocessing is always necessary as it enhances blurred or distorted images. Depending on the application exact filters should be applied and for this paper, median filter is chosen as the optimum one for noise removal.

ISSN (Online): 1694-0784
ISSN (Print): 1694-0814

Fig.3.a) Input Frame

b) Preprocessed frame

Fig.5 Region Growing Segmented Frame

## 2.3. Region Growing Segmentation

Effective Segmentation is carried out by using Seeded Region-growing (SRG) Segmentation algorithm. The image is converted into gray level first. The Seed values and the threshold values are provided. Region-growing approaches exploit the important fact that pixels, which are close together, have similar gray values.

***Algorithm 1*** *– Seeded Region Growing Algorithm*

Region growing procedure group pixels or sub-regions into larger regions based on predefined criteria of growth. Start with a single pixel (seed) and add new pixels slowly.

INPUT: Frame sequences
OUTPUT: Segmented Regions

1) Choose the seed pixel
2) Check the neighboring pixels and add them to the region if they are similar to the seed
3) Repeat step 2 for each of the newly added pixels; stop if no more pixels can be added.
4) More than one seeds can be used to segment the image.

Fig.4 Foreground Segmented frame

By selecting appropriate seed value and by setting a proper Threshold, the frames are segmented using Seeded Region Growing Segmentation algorithm. The popping crease at both the batsman and bowlers end is to be detected because the occurrence of the ball at these points is very high. This is done by horizontal line detection method. A mask H=[-1 -1 -1; 2 2 2;-1 -1 -1], which is nothing but a horizontal mask is applied to detect the horizontal crease lines in the input frame. The players, stumps, crease were identified because in these areas, the ball occurrence is more.

## 2.4. Ball Candidate Generation

The major idea behind this strategy is that while it is very challenging to achieve high accuracy when locating the precise location of the ball, it is relatively easy to achieve very high accuracy in locating the ball among a set of ball-like candidates.

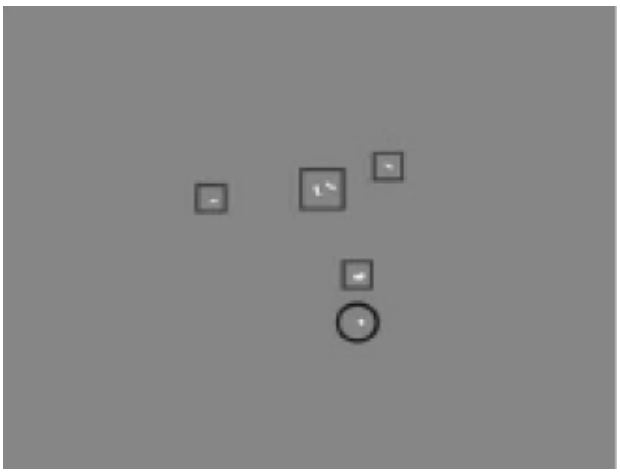


Fig.6 Sample Ball-candidates in a Sports video
(Image Courtesy: Xinguo Yu, H.W.Leong,
Changsheng Xu, Qi Tian)

ISSN (Online): 1694-0784
ISSN (Print): 1694-0814

The key challenges for ball selection [1] are:

1) There are many ball-like objects in a frame and
2) There is no universal ball representation that can be used to distinguish the ball from other ball-like objects in the frame.

To partially resolve this challenge, the first key idea is to focus on generating a set of ball-candidates for each frame instead of attempting to identify the ball in each frame, an approach named anti-model approach is used to remove non-ball candidates and generate only ball-candidates. The effectiveness of the anti-model approach is dependent on the accuracy of the sieves defined. By storing only a small set of ball-candidates per frame, we can process all the ball-candidates in a long sequence of frames at the same time. When the ball-candidates are processed together, rich spatial and temporal information can be obtained. Even after the ball-candidate generation, it is very hard to identify the ball from among the ball-candidates in since they have all the properties of the ball. However, the probability that the ball among the ball candidates is very high rather than its occurrence in the entire frame. This is used to detect and subsequently track the ball in the frame. The anti-model approach, which is a type of model based algorithm is accurate than feature based or motion based algorithm.

We can significantly reduce the rate of false negatives, but at the price of a temporarily higher rate of false positives. We can measure each ball-candidate against a number of properties of the "ideal" ball image and use these to classify the ball-candidates obtained. Several sieves such as shape, size, and color are used to sieve out the non-ball objects and the remaining objects are referred as ball-candidates which satisfy all the properties defined by the sieves. The properties used in the algorithm are size, and circularity. These are called sieves. Sieves Definition and pruning out the non-ball candidates forms the major part of the process.

The more the number of sieves, more accurate will be the ball candidates generated The Resulting output contains only the ball candidates.

***Algorithm 2*** *– Ball Candidate generation algorithm*

1) For each frame F, the set of objects in the frame are identified, denoted by N(F).
2) Then, a set of sieves were developed based on properties of the image of the ball in F and the sieves are used to sieve out non-ball objects in F.
3) The remaining objects are the ball-candidates and they satisfy all the properties defined by the sieves.
4) This remaining set are denoted by B(F) , the set of ball-candidates for frame .
5) The probability that the ball is among the ball-candidates, namely, b € B (F) is very high.

This can be used widely for other sports videos like golf, basket ball, volley ball, table tennis etc. For sports like table tennis, we need to reduce false alarms. So, this method can be used for ball detection. The ball detection accuracy will be tremendously increased if this method is adopted. This is fairly superior to other algorithms like Atherton algorithm, Modified Atherton algorithm, CHT algorithm for ball detection and tracking.

The objects similar to ball such as front view of a player's shoe, heads of players and umpire, bottle lying near the boundary boards and several others comes under Ball-candidates category. These alone are preserved in the frames and other objects in the frame are eliminated. This approach makes the task simpler. Instead of processing the entire frame to search the ball, it is better to search the ball among the ball-candidates which will reduce time consumption and complexity.

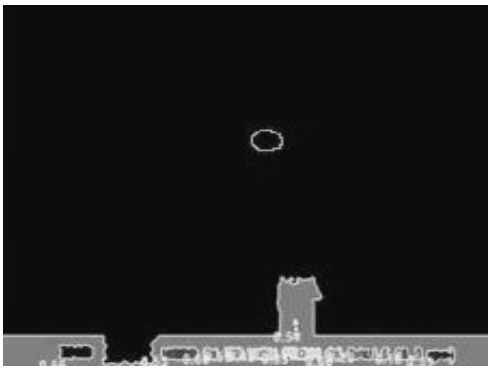


Fig.7 Ball and ball-candidates

The ball-candidates similar to ball are not present in the active region of the frame. They are only available as numbers in score board and hence can be neglected. To improve accuracy, neural networks is used to train the ball and from the Ball-candidates, the ball can be easily found among the Ball-candidates

## 3. Results and Discussion

The Results shown in the figures clearly depict the effectiveness of the ball detection algorithm for cricket. The Circular Hough Transform (CHT) based ball detection performs poorly when there are occlusions of the ball with the surface or with the players. The other proposed algorithms also did not give more accuracy in ball detection. We tested our system on real image sequences of actual Cricket video. India Vs Pakistan match held in Lahore was taken as input. The Frames were grabbed at a rate of 30 frames per Second at a resolution 320 X 240. Around 300 frames

ISSN (Online): 1694-0784
ISSN (Print): 1694-0814

were obtained for experimentation. The algorithm minimizes the false alarms to a great extent.

The second sets of tests were performed on a match between India Vs Sri Lanka held in Kandy. Around 250 frames were extracted at a rate of 30 frames per second at a resolution of 320 X 240. The performance was equally good as the previous video.

Several other tests were performed on the videos from 1996 Cricket World Cup matches and performances were analyzed. Satisfactory outputs were obtained with all the videos.

Moving Object Segmentation is a major problem when it comes to object detection in videos. This is resolved in this method in a simple manner. In each and every frame in these videos, about five to seven ball-like objects(ball-candidates) were obtained and all other objects got eliminated. From these objects, by further classification in a detailed manner (i.e.) by defining more and more sieves, ball alone can be detected from the ball-candidates.

Lot of methods like feature based algorithms, which uses certain features alone for ball detection, motion based methods which uses motion vectors to find the ball position in the subsequent frames were previously used for ball detection . But they are not as accurate as expected which compels for a novel method for ball detection and subsequent tracking.

## 4. Conclusion and Future Enhancements

This work has presented a novel approach to detect the Cricket ball from Cricket video images. To reach the goal, the non-ball objects are removed by shape, size, and candidate feature images are created. This approach can detect occluded balls and balls that are merged with other objects in the frame.

In future this can be applied to other detection-and-tracking problems, most important problems such as wildlife tracking and various surveillance tracking problems. Another direction is to use the approach for other higher level tasks such as event detection in the sports video. For soccer it can be applied for events such as detection of events such as kicking, passing, shooting, and team ball possession. For cricket events such as bowling, playing a stroke, fielding the ball etc, the events which may be difficult to analyze based only on the low-level feature.

The ball detection can be applied for all the sports videos for event detection. Ball detection in golf will be much more challenging than cricket because of the small ball size and distortions due to high velocity of the ball. Automatic Highlights extraction is another application which can be developed based on this work. This will provide backbone to highlights on demand service in cable TV networks. Shadow removal process can be achieved with a more robust algorithm. This case will both improve object silhouettes and detection results.

## References

[1] Xinguo Yu, Hon Wai Leong and Changsheng Xu, “Trajectory-based Ball Detection and tracking in Broadcast Soccer Video”, IEEE Transactions On Multimedia, Dec. 2006 pp 1164-1178.

[2] T. D’Orazio, C. Guarangnella, M. Leo, and A. Distante, “A new algorithm for ball recognition using circle Hough transform and neural classifier,” Pattern Recognition., vol. 37, pp. 393–408, 2004

[3] Xinguo Yu, Qi Tian, and Kong Wah Wan, “A Novel ball Detection Framework for Real Soccer Video” ,Proceedings of ICME 2003, 2003, pp 265-268.

[4] T. D’Orazio, N. Ancona, G. Cicirelli, and M. Nitti, “A ball detection algorithm for real soccer image sequences”, Proceedings of ICPR, 2002, pp 210-213

[5] Y. Ohno, J. Miura, and Y. Shirai, “Tracking players and a ball in soccer games” Proceedings of the International Conference on Multisensor Fusion and Integration for Intelligent Systems, Taipei, Taiwan, Aug. 1999.

[6] Y. Seo, S. Choi, H. Kim, and K. Hong, “Where are the ball and players? Soccer game analysis with color based tracking and image mosaic”, Proceedings of. ICIAP, Sep. 17–19, 1997, pp 196–203.

[7] P. Bouthemy and E. Francois, “Motion segmentation and qualitative scene analysis from an image sequence,” International journal on computer vision., vol. 10,pp. 157–182, 1993.

[8] Z. Zhang and O. D. Faugeras, “Three-dimensional motion computation and object segmentation in a long sequence of stereo frames,” International Journal on computer Vision, vol. 7, no. 3, pp. 211–241, 1992.

[9] K. P. Karmann, A. V. Brandt, and R. Gerl, “Moving object segmentation based on adaptive reference images,” in Proceedings of Conference. Eusipco, Barcelona, Spain, Sep. 1990, pp. 951–954.

[10] H. K. Yuen, J. Illingworth, and J. Kittler, “Detecting partially occluded ellipses using the Hough transform,” Image Vision Computing., vol. 7, pp.31–37, 1989.

[11] Z. Zhang and O. D. Faugeras, “Three-dimensional motion computation and object segmentation in a long sequence of stereo frames,” International journal on computer vision, vol. 7, no. 3, pp. 211–241, 1992.

[12] X. Yu, C. Xu, H. W. Leong, Q. Tian, Q. Tang, and K. W. Wan, “Trajectory-based ball detection and tracking with applications to semantic analysis of broadcast soccer video,” in Proceedings of ACM Conference on Multimedia,2003, pp. 11–20.

ISSN (Online): 1694-0784
ISSN (Print): 1694-0814

[13] T. Zhao and R. Nevatia, “Car detection in low resolution aerial image,” in Proceedings of ICCV, 2001, vol 1, pp. 710–717.

[14] Y. Gong, T. S. Lim, H. C. Chua, H. J. Zhang, and M. Sakauchi, “Automatic parsing of TV soccer programs,” in Proc. 2nd International. Conference on Multimedia Computers and Systems, 1995, pp. 167–174.

[15] Anil K. Jain, “Fundamentals of Digital Image Processing”, Pearson Education, second Indian reprint, 2004.

**B.L.Velammal** recieved Bachelors and Masters Degrees in 1999 and 2002 respectively, from Manonmaniam Sundaranar University. The author has worked as a Lecturer in various Engineering colleges. Currently, she is working as a Lecturer in Anna University. Some of the research works are published in International Journals. Her research area includes multimedia content adaptation and analysis.

**Anandhakumar P** received Ph.D degree in CSE from Anna University, in 2006. He is working as Assistant Professor in Dept. of IT, MIT Campus, Anna University. His research area includes image processing and networks. He has published a number of research works in International Journals. He is interested in guiding exceptional research works in the area of multimedia and networks.

# Modeling and Simulation of Microcode-based Built-In Self Test for Multi-Operation Memory Test Algorithms

**Dr. R.K. Sharma Aditi Sood**
**Department of Electronics and Communications Engineering**
**National Institute of Technology, Kurukshetra**

**Abstract**
As embedded memory area on-chip is increasing and memory density is growing, newer test algorithms like March SS are defined to detect newly developing faults. These new March algorithms contain multiple operations per March element. This paper presents a microcoded BIST architecture which can implement these new March tests having number of operations per element according to the growing needs of embedded memory testing. This is shown by implementing March SS Test and testing for new faults including Write Disturb Fault (WDF), Transition Coupling Fault (Cft), Deceptive Read Disturb Coupling Fault (Cfdrd), which established tests like March C- are not capable of detecting. Verilog HDL code of this architecture is written and synthesized using Xilinx ISE 8.2i. Verification of the architecture is done by testing Mentor's ModelSim.

***Keywords-*** *Defect-Per Million (DPM;, Built-In Self Test (BIST); Memory Built-in Self Test (MBIST;, Microcoded MBIST; MUT (Memory Under Test.)*

## 1. Introduction

According to the 2001 ITRS, today's system on chips (SoCs) are moving from logic dominant chips to memory dominant chips in order to deal with today's and future application requirements. The dominating logic (about 64% in 1999) is changing to dominating memory (approaching 90% by 2011) [1] as shown in Fig.1.

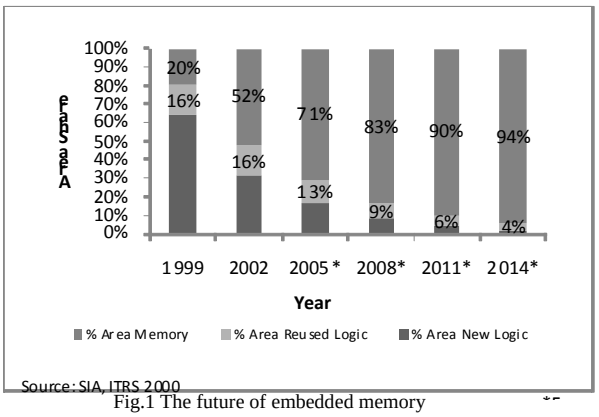


Fig.1 The future of embedded memory

As the memories grow in size and speed, the bit lines, word lines and address decoder pre-select lines will have high parasitic capacitance in addition to a high load. This increases their sensitivity for delay and timing related faults. Also, the significance of the resistive opens is considered to increase in current and future technologies. Since the partial resistive opens behave as delay and time related faults, these faults will become more important in the deep-submicron technologies [2]. Moreover, transistor short channel effect, cross talk effects, impact of process variation have to be necessarily taken into account for developing fault models for embedded memories based on newer technologies.

These factors help in the development of new, optimal, high coverage tests and diagnostic algorithms that allow for dealing with the new defects. The greater the fault detection and localization coverage, the higher the repair efficiency; hence higher the obtained yield.

Thus, the new trends in Memory testing will be driven by the following items:
• Fault modeling: New fault models should be established in order to deal with the new defects introduced by current and future (deep-submicron) technologies.
• Test algorithm design: Optimal test/diagnosis algorithms to guarantee high defect coverage for the new memory technologies and reduce the DPM level.
• BIST: The only solution that allows at-speed testing for embedded memories.

A new microcoded BIST architecture is presented here which is capable of employing new test algorithms like March SS [5] and March RAW [3] that have been developed for coverage of some recently developed static and dynamic fault models.

## 2. Microcode MBIST Controller

As shown in the previous section, the importance of developing new fault models increases with the new memory technologies.

The well-known fault models, developed before late 1990's could not explain the occurrence of many faults that were detected using experimental results based on DPM screening of a large number of tests applied to a large number of memory chips that were performed at that time, suggesting the existence of additional faults. This implied that new memory technologies involving high density of shrinking devices lead to newer faults and stimulated the introduction of new fault models, based on defect injection

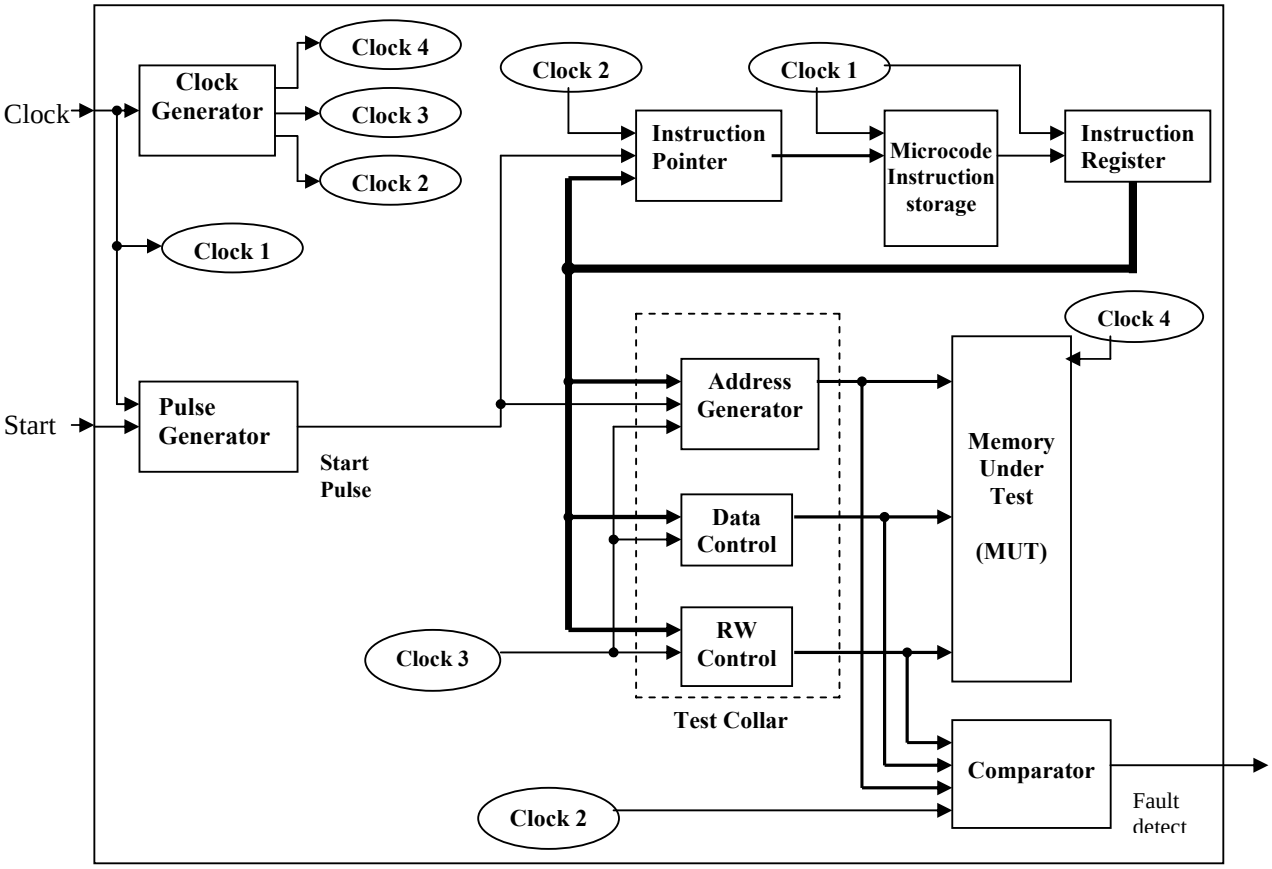



Fig. 2 Microcode MBIST Architecture

and SPICE simulation. Write Disturb Fault (WDF), Transition Coupling Fault (Cft), Deceptive Read Disturb Coupling Fault (Cfdrd) etc. are examples of some such newly defined fault models [2]. Another class of faults called Dynamic faults which require more than one operation to be performed sequentially in time in order to be sensitized have also been defined. [3-4]

Traditional tests, like March C-, are thus becoming insufficient/inadequate for today's and the future high speed memories. Therefore, more appropriate test algorithms have been developed to deal with these new fault models. Examples of such tests are March SS [5] and March RAW [3]. March SS covers some of the new fault models like Deceptive Read Destructive fault, Write disturb fault, etc., whereas March RAW covers some of the Dynamic faults.

These new test algorithms have as many as six or seven operations per march element, and thus some of the recently modeled and simulated architectures are inadequate to implement these test algorithms, as they have been developed to make space for only up to two test operations per March element [6]. This architecture is capable of implementing the newly developed March algorithms, because of its ability to execute algorithms with unlimited number of operations per March element. Thus many of the recently developed March algorithms can be applied using this architecture.

This has been illustrated in the present work by implementing March SS algorithm. However, the same hardware can be used to implement other new March algorithms also by just changing the Instruction storage unit, or the instruction codes and sequence inside the instruction storage unit. The instruction storage unit is used to store predetermined test pattern.

## 2.1 Methodology

The block diagram of the architecture is shown in Fig 2. The BIST Control Circuitry consists of Clock Generator, Pulse Generator, Instruction Pointer, Microcode Instruction storage unit, Instruction Register. The Test Collar circuitry consists of Address Generator, RW Control, Data Control.

**Clock Generator** generates simulated clock waveforms Clock2, Clock3, Clock4, for the rest of the circuitry based on the input clock (named Clock1) as shown in Fig. 3

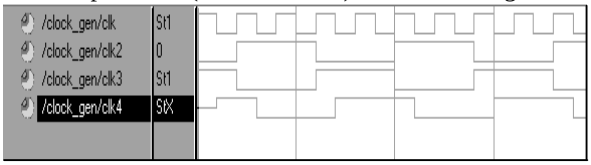



Fig 3. Simulated waveform of Clock generator Module

**Pulse Generator** generates a 'Start Pulse' at positive edge of the 'Start' signal which marks start of test cycle. **Instruction Pointer** points to the next microword, that is

the next march operation to be applied to the memory under test (MUT). Depending on the test algorithm, it is able to i) point at the same address, ii) point to the next address, or iii) jump back to a previous address.

The flowchart in Fig. 4 precisely describes the functioning of the Instruction Pointer. Here, 'Run complete' indicates that a particular march test operation has marched through the entire address space of MUT in increasing or decreasing order as dictated by the microcode.

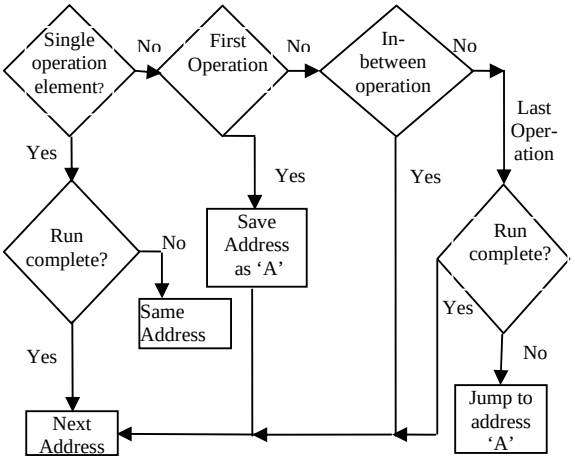



Fig. 4 Flowchart illustrating functional operation of Instruction Pointer

**Instruction Register (IR)** holds the microword (containing the test operation to be applied) pointed at by the Instruction Pointer. The various relevant bits of microword are sent to other blocks from IR.

**Address Generator** points to the next memory address in MUT, according to the test pattern sequence. It can address the memory in forwards as well as backwards direction.
**RW Control** generates read or write control signal for MUT, depending on relevant microword bits..
**Data Control** generates data to be written to or expected to be read out from the memory location being pointed at by the Address Generator

The Address Generator, RW Control and Data Control together constitute the Memory *Test Collar Comparator* gives the fault waveform which consists of positive pulses whenever the value being read out of the memory does not match the expected value as given by Test Collar.

## 2.2. Microcode Instruction specification.

The microcode is a binary code that consists of a fixed number of bits, each bit specifying a particular data or operation value. As there is no standard in developing a microcode MBIST instruction [7], the microcode instruction fields can be structured by the designer depending on the test pattern algorithm to be used.
The microcode instruction developed in this work is coded to denote one operation in a single microword. Thus a five operation March element is made up by five micro-code words. The format of 7-bit microcode MBIST instruction word is as shown in Fig. 5. Its various fields are explained as follows: Bit #1 (=1) indicates a valid microcode instruction, otherwise, it indicates the end of test for BIST Controller. Bits #2, #3 and #4 stand for first operation, in-between operation and last operation of a multi-operation March element, interpreted as shown in Fig. 5.
Bit #5 (=1) notifies that the memory under test (MUT) is to be addressed in decreasing order; else it is accessed in increasing order. Bit #6 (=1) indicates that the test pattern data is to be written into the MUT; else, it is retrieved

| #1 | #2 | #3 | #4 | #5 | #6 | #7 |
|---|---|---|---|---|---|---|
| Valid | Fo | Io | Lo | I/D | R/W | Data |

| Fo | Io | Lo | Description |
|---|---|---|---|
| 0 | 0 | 0 | A single operation element |
| 1 | 0 | 0 | First operation of a Multi-operation element |
| 0 | 1 | 0 | In-between Operation of a Multi-operation element |
| 0 | 0 | 1 | Last Operation of a Multi-operation element |

Fig. 5 Format of Microcode Instruction word

from the memory under test. Bit #7 (=1) signifies that a byte of 1s is to be generated (written to MUT or expected to be read out from the MUT); else byte containing all zeroes are generated.

Table 1 Content of Instruction Storage Unit

| | #1 Valid | #2 Fo | #3 Io | #4 Lo | #5 I/D (0/1) | #6 R/W (0/1) | #7 Data (0/1) |
|---|---|---|---|---|---|---|---|
| M0: χ W0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| M1: ↑{ R0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| R0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| W0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 |
| R1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| W1} | 1 | 0 | 0 | 1 | 0 | 1 | 1 |
| M2: ↑ {R1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 |
| R1 | 1 | 0 | 1 | 0 | 0 | 0 | 1 |
| W1 | 1 | 0 | 1 | 0 | 0 | 1 | 1 |
| R1 | 1 | 0 | 1 | 0 | 0 | 0 | 1 |
| W0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 |
| M3: ↓{R0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 |
| R0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 |
| W0 | 1 | 0 | 1 | 0 | 1 | 1 | 0 |
| R0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 |
| W1} | 1 | 0 | 0 | 1 | 1 | 1 | 1 |
| M4: ↓{ R1 | 1 | 1 | 0 | 0 | 1 | 0 | 1 |
| R1 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| W1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 |
| R1 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| W0} | 1 | 0 | 0 | 1 | 1 | 1 | 0 |
| M5: χ R0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |
| | 0 | X | X | X | X | X | X |

The instruction word is so designed so as to represent any March algorithm. The contents of Instruction storage unit for March SS algorithm are shown in Table 1.

The first march element M0 is a single operation element, which writes zero to all memory cells in any order. Similarly, the second march element M1 is a multi-operation element, which consists of five operations: i) R0, ii) R0, iii) W0, iv) R1 and v) W1. MUT is addressed in increasing order as each of these five operations is performed on each memory location before moving on to the next location.

### 2.3 Behavior Simulation

Mentor's ModelSim has been used to verify the functionality and timing constraints of BIST module.
Verilog HDL code of the above architecture written and synthesized using Xilinx ISE 8.2i [8].

## 3. RESULTS

The simulation waveform of a fault-free SRAM is shown in Fig. 6. Of the generated Clock signals, only Clock2 and Clock4 appear in the top module

The top module shows the interfacing of BIST Controller (including test collar), MUT and Comparator. As the START signal goes high, indicating the start of test, the first March element M0 of March SS algorithm is executed. As this is a write signal, no values are read out from the memory to be compared with expected or correct values and hence the output FAULT waveform of comparator is high impedance. As read operation starts at the beginning of execution of M1 element, the values from MUT are read out and compared with the expected values. The FAULT waveform shows a 'low' level throughout the test for a fault-free SRAM

The SRAM model is also amended to be in defective state by inserting faults. The simulated waveform is shown in Fig. 7.

The inserted faults are Deceptive Read Disturb fault (DRDF) at location 11, Write Disturb Fault (WDF) at location 13, Deceptive Read Disturb Coupling fault (CFdrd) at location 9 (victim) due to location 10 (aggressor), Write Disturb Coupling Fault (CFwd) at location 14 (victim) due to location 15 (aggressor) [9].
The fault detect waveform shows 12 pulses due to the above faults in given four locations, as the test elements march through MUT to uncover these defects.
The above stated faults cannot be detected by March C- algorithm but are easily detected by March SS Algorithm which can be implemented by the architecture presented in this work.

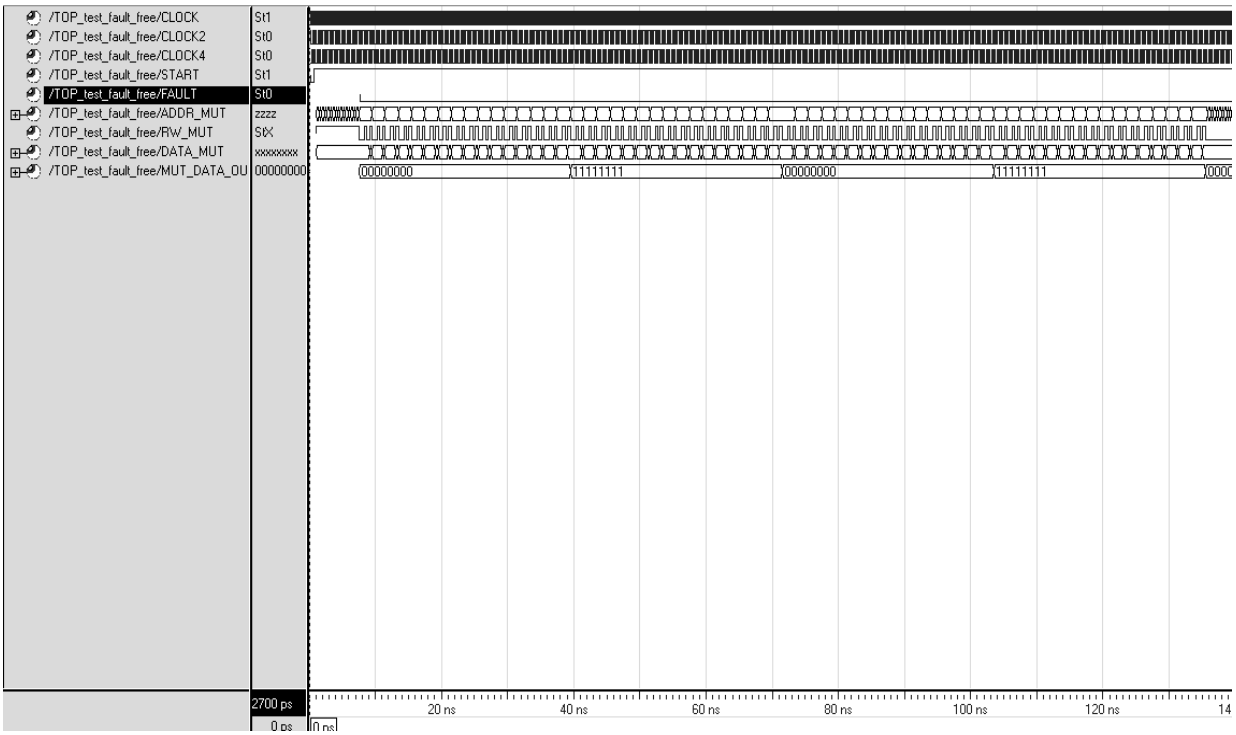



Fig. 6 Simulated waveform of fault-free SRAM

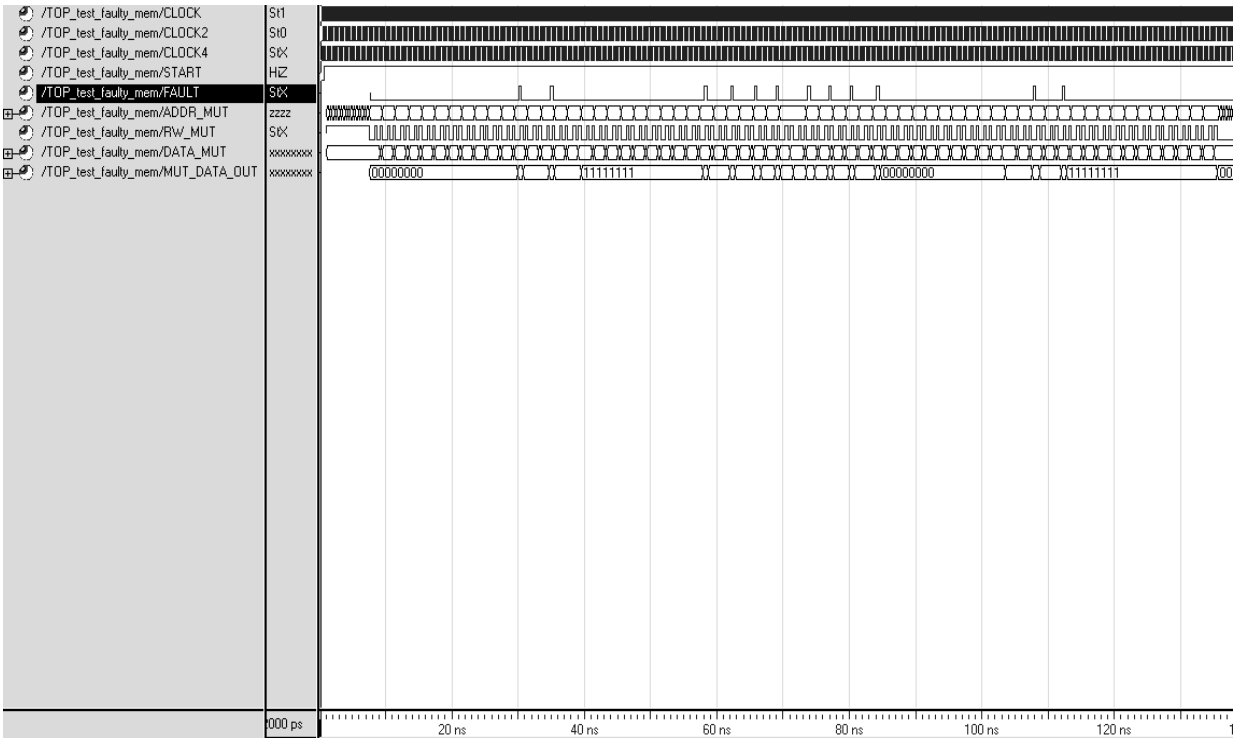



Fig. 7 Simulated waveform of faulty SRAM

## 4. Conclusion

The simulation results have shown that the micro-coded MBIST architecture described here is an effective testing method to test embedded memories as it provides a flexible approach and better fault coverage. Just as March SS, any new march algorithm can be implemented using the same BIST hardware by changing the instructions in the microcode storage unit, without the need to redesign the entire circuitry.

## References

[1] International SEMATECH, "International Technology Roadmap for Semiconductors (ITRS): Edition 2001"

[2] S. Hamdioui, G.N. Gaydadjiev, A.J .van de Goor, "State-of-art and Future Trends in Testing Embedded Memories", *International Workshop on Memory Technology, Design and Testing (MTDT'04)*, 2004.

[3] S. Hamdioui, Z. Al-Ars, A.J. van de Goor, "Testing Static and Dynamic Faults in Random Access Memories", *In Proc. of IEEE VLSI Test Symposium*, pp. 395-400, 2002.

[4] S. Hamdioui, et. al, "Importance of Dynamic Faults for New SRAM Technologies", *In IEEE Proc. Of European Test Workshop*, pp. 29-34, 2003.

[5] S. Hamdioui, A.J. van de Goor and M. Rodgers, "March SS: A Test for All Static Simple RAM Faults", *In Proc. of IEEE International Workshop on Memory Technology, Design, and Testing*, pp. 95-100, Bendor, France, 2002.

[6] N. Z. Haron, S.A.M. Junos, A.S.A. Aziz, "Modelling and Simulation of Microcode Built-In Self test Architecture for Embedded Memories", *In Proc. of IEEE International Symposium on Communications and Information Technologies* pp. 136-139, 2007.

[7] R. Dean Adams, "High Performance Memory Testing: Design Principles, Fault Modeling and Self-Test", Springer US, 2003.

[8] "Xilinx ISE 6 Software Manuals and help – PDF Collection",http://toolbox.xilinx.com/docsan/xilinx7/books/ manuals .pdf

[9] A.J. van de Goor and Z. Al-Ars, "Functional Fault Models: A Formal Notation and Taxonomy", *In Proc. of IEEE VLSI Test Symposium*, pp. 281-289, 2000.

[10] Zarrineh, K. and Upadhyaya, S.J., "On Programmable memory built-in self test architectures," *Design, Automation and Test in Europe Conference and Exhibition 1999*. Proceedings , 1999, pp. 708 -713

[11] Sungju Park et al, "Microcode-Based Memory BIST Implementing Modified March Algorithms", *Journal of the Korean Physical Society*, Vol. 40, No. 4, April 2002, pp. 749-753

[12] A.J. van de Goor, "Using March tests to test SRAMs", *Design & Test of Computers, IEEE*, Volume: 10, Issue: 1, March 1993 Pages: 8-14.

[13] R. Dekker, F. Beenker and L. Thijssen, "*Fault Modeling and Test Algorithm Development for Static Random Access Memories*",

[14] R.Dekker, F. Beenker, L. Thijssen. "A realistic fault model and test algorithm for static random access memories". *IEEE Transactions on CAD*, Vol. 9(6), pp 567-572, June 1990.

[15] B. F. Cockburn: "Tutorial on Semiconductor Memory Testing" *Journal of Electronic Testing: Theory and Applications*, 5, pp 321-336 1994 Kluwer Academic Publishers, Boston.

[16] A.J. van de Goor, "Testing Semiconductor Memories, Theory and Practice" ComTex Publishing, Gouda, Netherlands, 1998.

IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 3, No 2, May 2010
ISSN (Online): 1694-0784
ISSN (Print): 1694-0814


# A Descriptive Classification of Causes of Data Quality Problems in Data Warehousing

Ranjit Singh[1], Dr. Kawaljeet Singh[2]

[1] Research Scholar, University College of Engineering (UCoE), Punjabi University
Patiala (Punjab), INDIA

[2] Director, University Computer Center (UCC), Punjabi University
Patiala (Punjab), INDIA

**Abstract**

Data warehousing is gaining in eminence as organizations become awake of the benefits of decision oriented and business intelligence oriented data bases. However, there is one key stumbling block to the rapid development and implementation of quality data warehouses, specifically that of warehouse data quality issues at various stages of data warehousing. Specifically, problems arise in populating a warehouse with quality data. Over the period of time many researchers have contributed to the data quality issues, but no research has collectively gathered all the causes of data quality problems at all the phases of data warehousing Viz. 1) data sources, 2) data integration & data profiling, 3) Data staging and ETL, 4) data warehouse modeling & schema design. The state-of-the-art purpose of the paper is to identify the reasons for data deficiencies, non-availability or reach ability problems at all the aforementioned stages of data warehousing and to formulate descriptive classification of these causes. We have identified possible set of causes of data quality issues from the extensive literature review and with consultation of the data warehouse practitioners working in renowned IT giants on India. We hope this will help developers & Implementers of warehouse to examine and analyze these issues before moving ahead for data integration and data warehouse solutions for quality decision oriented and business intelligence oriented applications.

**Keywords : Data Quality (DQ), ETL, Data Staging, Data Warehouse**

## Section I – Introduction

### 1.1 Understanding Data Quality

The existence of data alone does not ensure that all the management functions and decisions can be smoothly undertaken. The one definition of data quality is that it's about bad data - data that is missing or incorrect or invalid in some context. A broader definition is that data quality is achieved when organization uses data that is comprehensive, understandable, consistent, relevant and timely. Understanding the key data quality dimensions is the first step to data quality improvement. To be process able and interpretable in an effective and efficient manner, data has to satisfy a set of quality criteria. Data satisfying those quality criteria is said to be of high quality. Abundant attempts have been made to define data quality and to identify its dimensions. Dimensions of data quality typically include accuracy, reliability, importance, consistency, precision, timeliness, fineness, understandability, conciseness and usefulness. For our research paper we have under taken the quality criteria by taking 6 key dimensions as depicted below figure1.

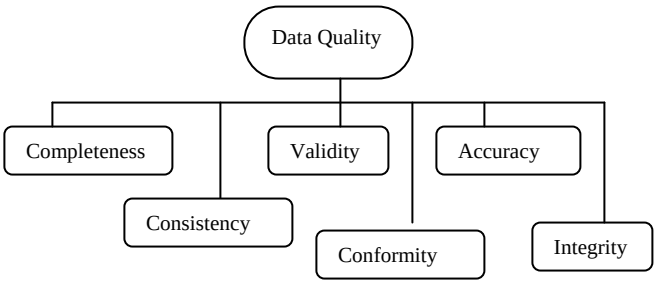



Figure 1: Data Quality Criteria [21]

***Completeness:*** deals with to ensure is all the requisite information available? Are some data values missing, or in an unusable state?

***Consistency:*** Do distinct occurrences of the same data instances agree with each other or provide conflicting information. Are values consistent across data sets?

***Validity:*** refers to the correctness and reasonableness of data

***Conformity:*** Are there expectations that data values conform to specified formats? If so, do all the values

ISSN (Online): 1694-0784
ISSN (Print): 1694-0814

conform to those formats? Maintaining conformance to specific formats is important.

***Accuracy:*** Do data objects accurately represent the "real-world" values they are expected to model? Incorrect spellings of product or person names, addresses, and even untimely or not current data can impact operational and analytical applications.

***Integrity:*** What data is missing important relationship linkages? The inability to link related records together may actually introduce duplication across your systems.

### 1.2 Data Warehousing

Data warehouses are one of the foundations of the Decision Support Systems of many IS operations. As defined by the "father of data warehouse", William H. Inmon, a data warehouse is "a collection of Integrated, Subject-Oriented, Non Volatile and Time Variant databases where each unit of data is specific to some period of time. Data Warehouses can contain detailed data, lightly summarized data and highly summarized data, all formatted for analysis and decision support" (Inmon, 1996). In the "Data Warehouse Toolkit", Ralph Kimball gives a more concise definition: "a copy of transaction data specifically structured for query and analysis" (Kimball, 1998). Both definitions stress the data warehouse's analysis focus, and highlight the historical nature of the data found in a data warehouse.

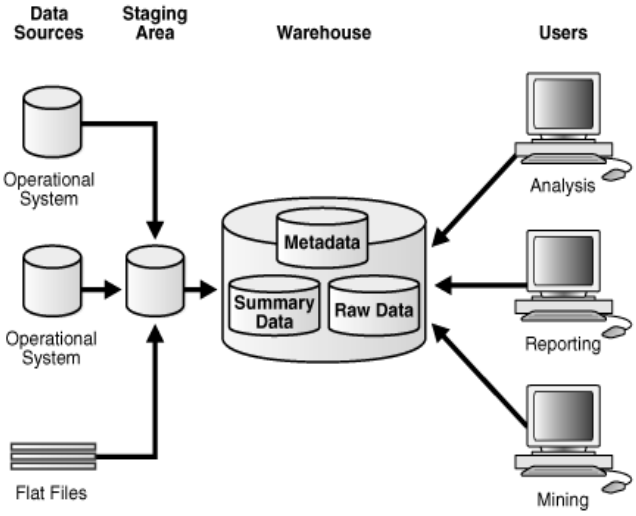



Figure 2: Data Warehousing Structure

### 1.3 Stages of Data Warehousing Susceptible to Data Quality Problems

The purpose of paper here is to formulate a descriptive taxonomy of all the issues at all the stages of Data Warehousing. The phases are:

- Data Source
- Data Integration and Data Profiling
- Data Staging and ETL
- Database Scheme (Modeling)

Quality of data can be compromised depending upon how data is received, entered, integrated, maintained, processed (Extracted, Transformed and Cleansed) and loaded. Data is impacted by numerous processes that bring data into your data environment, most of which affect its quality to some extent. All these phases of data warehousing are responsible for data quality in the data warehouse. Despite all the efforts, there still exists a certain percentage of dirty data. This residual dirty data should be reported, stating the reasons for the failure in data cleansing for the same.

Data quality problems can occur in many different ways [9]. The most common include:

- Poor data handling procedures and processes.
- Failure to stick on to data entry and maintenance procedures.
- Errors in the migration process from one system to another.
- External and third-party data that may not fit with your company data standards or may otherwise be of unconvinced quality.

The assumptions undertaken are that data quality issues can arise at any stage of data warehousing viz. in data sources, in data integration & profiling, in data staging, in ETL and database modeling. Following model is depicting the possible stages which are vulnerable of getting data quality problems

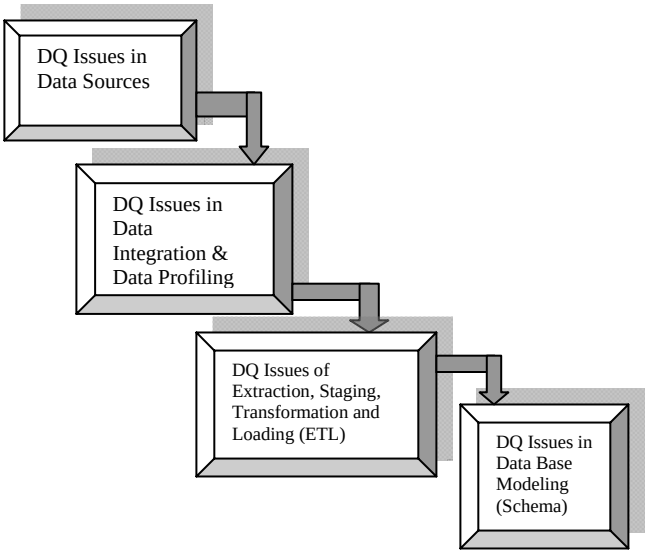



Figure 3: Stages of Data Warehouse Susceptible for DQ Problems

ISSN (Online): 1694-0784
ISSN (Print): 1694-0814

## Section II

### 2.1 Methodology

The study is designed as a literature review of materials published between 1992 and 2008 on the topics of data quality and data warehouses. The Figure3 presents the resulting research model formulated trough extensive literature review. To develop the research model, the IT implementations infrastructure, data warehousing literature, research questionnaires related to data quality were reviewed to identify various reasons of data duality problems at the stages mentioned in the model. Classification of causes of data quality problems so formed will be divided into the factors responsible for data quality at the phases. Later in the next phase of study, it will be converted into survey instrument for the confirmation of these issues from the data warehouse practitioners.

#### 2.1.1 Literature Reviewed.

***Channah E Naiman & Aris M. Ouksel (1995)-*** the paper proposed a classification of semantic conflicts and highlighted the issue of semantic heterogeneity, schema integration problems which further may have far reaching consequences on data quality

***John Hess (1998)*** the report has highlighted the importance of handling of missing values of the data sources, specially emphasized on missing dimension attribute values.

***Jaideep Srivastava, Ping-Yao Chen (1999)*** the principal goal of this paper is to identify the common issues in data integration and data-warehouse creation. Problems arise in populating a warehouse with existing data since it has various types of heterogeneity.

***Amit Rudra and Emilie Yeo (1999)*** the paper concluded that the quality of data in a data warehouse could be influenced by factors like: data not fully captured, heterogeneous system integration and lack of policy and planning from management.

***Scott W. Ambler (2001)*** the article explored the wide variety of problems with the legacy data, including data quality, data design, data architecture, and political/process related issues. The article has provided a brief bifurcation of common issues of legacy data which contribute to the data quality problems.

***Won Kim et al (2002)*** paper presented a comprehensive taxonomy of dirty data and explored the impact of dirty data on data mining results. A comprehensive classification of dirty data is developed for use as a framework for understanding how dirty data arise, manifest themselves, and may be cleansed to ensure proper construction of data warehouses and accurate data analysis.

***Wane Eckerson & Colin White (2003)*** report says ETL tools are the traffic cop for business intelligence applications. They control the flow of data between myriad source systems and BI applications. As BI environments expand and grow more complex, ETL tools need to change to keep pace. ETL tools need to evolve from batch-oriented, single-threaded processing that extracts and loads data in bulk to continuous, parallel processes that capture data in near real time. They also need to provide enhanced administration and ensure reliable operations and high availability.

***Wayne Eckerson (2004)*** data warehousing projects gloss over the all-important step of scrutinizing source data before designing data models and ETL mappings. The paper presented the reasons for data quality problems out of which most important are 1) Discovering Errors Too Late 2) Unreliable Meta Data. 3) Manual Profiling. 4) Lack of selection of automated profiling tools

### 2.2 Classification of Data Quality Issues

In order for the analyst to determine the scope of the underlying root causes of data quality issues and to plan the design the tools which can be used to address data quality issues, it is valuable to understand these common data quality issues. For the purpose of it the classification formed will be highly helpful to the data warehouse and data quality community.

#### 2.2.1 Data Quality Issues at Data Sources

A leading cause of data warehousing and business intelligence project failures is to obtain the wrong or poor quality data. Eventually data in the data warehouse is fed from various sources as depicted in the figure 4. The source system consists of all those 'transaction/Production' raw data providers, from where the details are pulled out for making it suitable for Data Warehousing. All these data sources are having their own methods of storing data. Some of the data sources are cooperative and some might be non cooperative sources. Because of this diversity several reasons are present which may contribute to data quality problems, if not properly taken care of. A source that offers any kind of unsecured access can become unreliable-and ultimately contributing to poor data quality.

Different data Sources have different kind of problems associated with it such as data from legacy data sources (e.g., mainframe-based COBOL programs) do not even have metadata that describe them. The sources of dirty data include data entry error by a human or computer system, data update error by a human or computer system. Part of the data comes from text files,

ISSN (Online): 1694-0784
ISSN (Print): 1694-0814

part from MS Excel files and some of the data is direct ODBC connection to the source database [16].

Some files are result of manual consolidation of multiple files as a result of which data quality might be compromised at any step. Table 1 summarizes the possible causes of data quality problems at data sources stage of data warehousing.

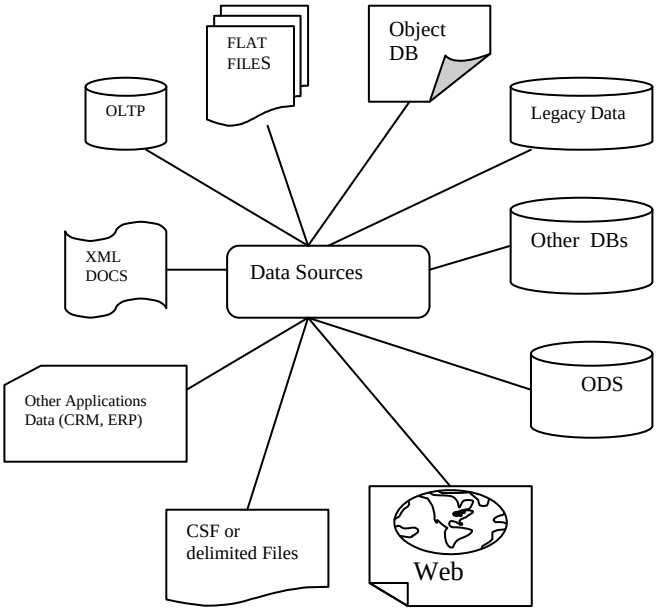



Fig 4: Possible Data Sources for Data Warehousing

Table 1:- Causes of Data Quality Problems at Data Sources Stage

| Sr. No | *CAUSES OF DATA QUALITY PROBLEMS AT DATA SOURCES* |
|---|---|
| 1 | *Inadequate selection of candidate data sources cause DQ Problems (sources which do not comply to business rules)* |
| 2 | *As time and proximity from the source increase, the chances for getting correct data decrease [4].* |
| 3 | *In adequate knowledge of inter dependencies among data sources incorporate DQ problems.* |
| 4 | *Inability to cope with ageing data contribute to data quality problems.[4]* |
| 5 | *Varying timeliness of data sources [6] [7].* |
| 6 | *Lack of validation routines at sources causes DQ Problems.* |
| 7 | *Unexpected changes in source systems cause DQ Problems.* |
| 8 | *Multiple data sources generate semantic heterogeneity which leads to data quality issues [1][4]* |
| 9 | *The complexity of a data warehouse increases geometrically with the span of time of data to be fed into it* |
| 10 | *Usage of decontrolled applications and databases as data sources for data warehouse in the organizations.* |
| 11 | *Use of different representation formats in data sources.* |
| 12 | *Measurement errors [11].* |
| 13 | *Non-Compliance of data in data sources with the Standards.* |
| 14 | *Failure to update the sources in a timely manner causes DQ Problems.* |
| 15 | *Failure to update all replicas of data causes DQ Problems.* |
| 16 | *Presence of duplicate records of same data in multiple sources cause DQ Problems [6] [7] [11].* |
| 17 | *Approximations or surrogates used in data sources.* |
| 18 | *Contradictory information present in data sources cause data quality problems [6] [7].* |
| 19 | *Different encoding formats (ASCII, EBCDIC,….) [11]* |
| 20 | *Inadequate data quality testing on individual data source lead to poor data quality.* |
| 21 | *Lack of business ownership, policy and planning of the entire enterprise data contribute to data quality problems. [4]* |
| 22 | *Columns .having incorrect data values (for Eg. The AgeInYear Column for a person contain value 3 although the birthdate column is having value Aug, 14, 1967) [6] [7].* |
| 23 | *Having Inconsistent/Incorrect data formatting (The name of a person is stored in one table in the format "Firstname Surname" and in another table in the forma t "Surname, Firstname") [6] [7] [11].* |
| 24 | *System fields designed to allow free forms (Field not having adequate length).* |
| 25 | *Missing Columns (You need a middle name of a person but a column for it does not exist.) [6][7].* |
| 26 | *Missing values in data sources [2][11][12].* |
| 27 | *Misspelled data [11][12]* |
| 28 | *Additional columns [6] [7] [11].* |
| 29 | *Multiple sources for the same data ((For Eg. customer information is stored in three separate legacy databases)[11].* |
| 30 | *Various key strategies for the same type of entity (for Eg. One table stores customer information using the Social Security Number as the key, another uses the ClientID as the key, and another uses a surrogate key) [6] [7].* |
| 31 | *Inconsistent use of special characters (for Eg. A date uses hyphens to separate the year, month, and day whereas a numerical value stored as a string* |

ISSN (Online): 1694-0784
ISSN (Print): 1694-0814

| | |
|---|---|
| | *uses hyphens to indicate negative numbers)[11] [20].* |
| 32 | *Different data types for similar columns (A customer ID is stored as a number in one table and a string in another).* |
| 33 | *Varying default values used for missing data [6] [7].* |
| 34 | *Various representations of data in source data (The day of the week is stored as T, Tues, 2, and Tuesday in four separate Columns) [6] [7] [20].* |
| 35 | *Lack of record level validation in source Data.* |
| 36 | *Data values stray from their field description and business rules (Such as the Maiden Name column is being used to store a person's Hobbies, zip code into phone number box) [6] [7].* |
| 37 | *Inappropriate data relationships among tables.* |
| 38 | *Unrealized data relationships between data members.* |
| 39 | *Not specifying NULL character properly in flat files data sources result in wrong data.* |
| 40 | *Delimiter that comes as a character in some field of the file may represent different meaning of data than the actual one.* |
| 41 | *Wrong number of delimiters in the sources (Files) causes DQ Problems.* |
| 42 | *Presence of Outliers.* |
| 43 | *Orphaned or dangling data (Data Pointing to other data which does not exists)[11]* |
| 44 | *Data and metadata mismatch.* |
| 45 | *Important entities, attributes and relationships are hidden and floating in the text fields [6][7].* |
| 46 | *Inconsistent use of special characters in various sources [6][7].* |
| 47 | *Multi-purpose fields present in data sources.* |
| 48 | *Deliberate Data entry errors (Input errors) contribute to data quality issues [4][11].* |
| 49 | *Poor data entry training causes data quality problems in data sources [11] [24].* |
| 50 | *Wrongly designed data entry forms, allowing illegibility [11] [24].* |
| *51* | *Different business rules of various data sources creates problem of data quality.* |
| *52* | *Insufficient plausibility (comparison within the data set and within time) checks in operative systems (e.g. during data input). [20]* |

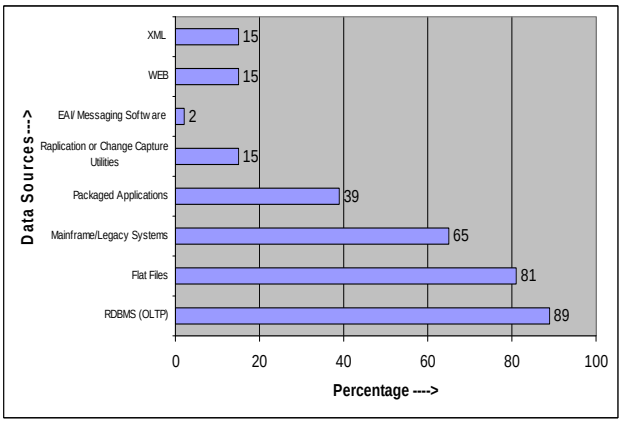
According to The Standish Group report, one of the primary causes of data migration project overruns and failures is a lack of understanding of the source data prior to data movement into some decision oriented data store such as data warehouse [17]. Especially the legacy sources are to be taken care of before we move data. So for the purpose of formulation of classification of causes of data pollution at the data source stage, we mainly identified data quality issues in following types of data sources

a) Legacy Systems
b) OLTP/ operational Systems
c) Flat/Delimited Files

And classification is confined to non multimedia (Images, Video and Audio) data only. The data sources considered is on the basis of survey conducted by Wayne Eckerson [14] in the report titled "Evaluating ETL and Data Integration Platforms". According to the survey, on average, organizations now extract data from 12 distinct data sources out which maximum is OLTP, Legacy and Flat Files by being encouraged from the report. Result of analysis is shown in Figure 5 showing the percentage of each type of data source from where companies extract data for the purpose of loading it into data warehouse.

All the causes presented in the table 1 are related to the data sources which are the feeder systems for the data warehouse. In literature problem of missing data, non standardized data and formats, and problems of data quality in legacy systems were emphasized more. One of the fundamental obstacles in the current data warehousing environment concerns the existence of inconsistent data. According to Amit Rudra and Emilie Yeo [4] as per their survey the top three reasons for data pollution in the data warehouse seems to be: Data never being fully captured, heterogeneous system integration and Lack of policy and planning from management. Our classification has presented much more number of causes of data pollution.

Figure 5:- Types of data sources Organizations extract data from. [14]




### 2.2.2 Causes of Data Quality Issues at Data Profiling Stage

When possible candidate data sources are identified and finalized data profiling comes in play immediately. Data profiling is the examination and

ISSN (Online): 1694-0784
ISSN (Print): 1694-0814

assessment of your source systems' data quality, integrity and consistency sometimes also called as source systems analysis. Data profiling is a fundamental, yet often ignored or given less attention as result of which data quality of the data warehouse is compromised At the beginning of a data warehouse project, as soon as a candidate data source is identified, a quick data profiling assessment should be made to provide a go/no-go decision about proceeding with the project. Table 2 is depicting the possible causes of data quality degradation at the data profiling stage of data warehousing.

Table 2: Causes of Data Quality Issues at Data Profiling Stage

| Sr. No | *CAUSES OF DATA QUALITY ISSUES AT DATA PROFILING.* |
|---|---|
| 1 | *Insufficient data profiling of data sources is responsible for data quality issues.* |
| 2 | *Manually derived information about the data Contents in operational systems propagates poor data quality [8].* |
| 3 | *Inappropriate selection of Automated profiling tool cause data quality issues [8].* |
| 4 | *Insufficient data content analysis against external reference data causes data quality problems.* |
| 5 | *Insufficient structural analysis of the data sources in the profiling stage.* |
| 6 | *Insufficient Pattern analysis for given fields within each data store.* |
| 7 | *Insufficient column profiling, single table structural profiling, cross table structural profiling of the data sources causes data quality problems [9].* |
| 8 | *Insufficient range and distribution of values or threshold analysis for required fields.* |
| 9 | *Lack of analysis of counts like record count, sum, mode, minimum, maximum percentiles, mean and standard deviation.* |
| 10 | *Undocumented, alterations identified during profiling cause data quality problems.* |
| 11 | *Inappropriate profiling of the formats, dependencies, and values of source data* |
| 12 | *Inappropriate parsing and standardization of records and fields to a common format* |
| 13 | *Lack of identification of missing data relationships* |
| 14 | *Hand coded data profiling is likely to be incomplete and leave the data quality problems.* |
| 15 | *Unreliable and incomplete metadata of the data sources cause data quality problems [8].* |
| 16 | *User Generated SQL queries for the data profiling purpose leave the data quality problems.* |
| 17 | *Inability of evaluation of inconsistent business processes during data profiling cause data quality problems.* |
| 18 | *Inability of evaluation of data structure, data values and data relationships before data integration, propagates poor data quality.* |
| *19* | *Inability of integration between data profiling, ETL cause no proper flow of metadata which leave data quality problems.* |

### 2.2.2 Data Quality issues at Data Staging ETL (Extraction, Transformation and Loading)

One consideration is whether data cleansing is most appropriate at the source system, during the ETL process, at the staging database, or within the data warehouse [15] [18]. A data cleaning process is executed in the data staging area in order to improve the accuracy of the data warehouse. The data staging area is the place where all 'grooming' is done on data after it is culled from the source systems. Staging and ETL phase is considered to be most crucial stage of data warehousing where maximum responsibility of data quality efforts resides. It is a prime location for validating data quality from source or auditing and tracking down data issues. There may be several reasons of data quality problems at this phase, some of the identified reasons from literature review are shown in Table 3.

Table 3: Causes of Data Quality Issues at Data Staging and ETL Phase

| Sr. No | *CAUSES OF DATA QUALITY ISSUES AT DATA STAGING AND ETL PHASE.* |
|---|---|
| 1 | *Data warehouse architecture undertaken affects the data quality (Staging, Non Staging Architecture).* |
| 2 | *Type of staging area, relational or non relational affects the data quality.* |
| 3 | *Different business rules of various data sources creates problem of data quality.* |
| 4 | *Business rules lack currency contributes to data quality problems [4].* |
| 5 | *The inability to schedule extracts by time, interval, or event cause data quality problems.* |
| 6 | *Lack of capturing only changes in source files [24].* |
| 7 | *Lack of periodical refreshing of the integrated data storage (Data Staging area) cause data quality degradation.* |
| 8 | *Truncating the data staging area cause data quality problems because we can't get the data back to reconcile.* |
| 9 | *Disabling data integrity constraints in data staging tables cause wrong data and relationships to be extracted and hence cause data quality problems [11].* |
| 10 | *Purging of data from the Data warehouse cause data quality problems [24].* |

ISSN (Online): 1694-0784
ISSN (Print): 1694-0814

| | |
|---|---|
| 11 | *Hand coded ETL tools used for data warehousing lack in generating single logical meta data store, which leads to poor data quality.* |
| 12 | *Lack of centralized metadata repository leads to poor data quality.* |
| 13 | *Lack of reflection of rules established for data cleaning, into the metadata causes poor data quality.* |
| 14 | *Inappropriate logical data map prepared cause data quality issues.* |
| 15 | *Misinterpreting/Wrong implementation of the slowly changing dimensions (SCD) strategy in ETL phase causes massive data quality problems.* |
| 16 | *Inconsistent interpretation or usage of codes symbols and formats [4].* |
| 17 | *Improper extraction of data to the required fields causes data quality problems [4].* |
| *18* | *Lack of proper functioning of the extraction logic for each source system (historical and incremental loads) cause data quality problems.* |
| *19* | *Unhandled null values in ETL process cause data quality problems.* |
| *20* | *Lack of generation of data flow and data lineage documentation by the ETL process causes data quality problems.* |
| *21* | *Lack of availability of automated unit testing facility in ETL tools cause data quality problems.* |
| *22* | *Lack of error reporting, validation, and metadata updates in ETL process cause data quality problems.* |
| *23* | *Inappropriate handling of rerun strategies during ETL causes data quality problems.* |
| *24* | *Inappropriate handling of audit columns such as created date, processed date and updated date in ETL* |
| *25* | *Inappropriate ETL process of update strategy (insert/update/delete) lead to data quality problems.* |
| *26* | *Type of load strategy opted (Bulk, batch load or simple load) cause Data Quality problems. [24]* |
| *27* | *Lack of considering business rules by the transformation logic cause data quality problems.* |
| *28* | *Non standardized naming conventions of the ETL processes (Jobs, sessions, Workflows) cause data quality problems.* |
| *29* | *Wrong impact analysis of change requests on ETL cause data quality problems.* |
| *30* | *Loss of data during the ETL process (rejected records) causes data quality problems. (refused data records in the ETL process)* |
| *31* | *Poor system conversions, migration, reengineering or consolidation contribute to the data quality problems [4] [24].* |
| *32* | *The inability to restart the ETL process from checkpoints without losing data [14]* |
| *33* | *Lack of Providing internal profiling or integration to third-party data profiling and cleansing tools.[14]* |
| *34* | *Lack of automatically generating rules for ETL tools to build mappings that detect and fix data defects[14]* |
| *35* | *Inability of integrating cleansing tasks into visual workflows and diagrams[14]* |
| *36* | *Inability of enabling profiling, cleansing and ETL tools to exchange data and meta data[14]* |

**2.2.4 Causes of Data Quality Problems at Data Modeling (Database Schema Design) Stage.**

The quality of the information depends on 3 things: (1) the quality of the data itself, (2) the quality of the application programs and (3) the quality of the database schema [19]. Design of the data warehouse greatly influences the quality of the analysis that is possible with data in it. So, special attention should be given to the issues of schema design. Some of the issues such as slowly changing dimensions, rapidly changing dimension, and multi valued dimensions etc. A flawed schema impacts negatively on information quality. Table 4 is depicting the listing of some most important causes of data quality issues at data warehouse schema designing

Table 4: Causes of Data Quality Issues at Data Warehouse Schema Modeling Phase

| Sr. No | *CAUSES OF DATA QUALITY ISSUES AT DATA WAREHOUSE SCHEMA DESIGN.* |
|---|---|
| 1 | *Incomplete or wrong requirement analysis of the project lead to poor schema design which further casue data quality problems.* |
| 2 | *Lack of currency in business rules cause poor requirement analysis which leads to poor schema design and contribute to data quality problems.* |
| 3 | *Choice of dimensional modeling (STAR, SNOWFLAKE, FACT CONSTALLATION) schema contribute to data quality.* |
| 4 | *Late identification of slowly changing dimensions contribute to data quality problems.* |
| 5 | *Late arriving dimensions cause DQ Problems.* |
| 6 | *Multi valued dimensions cause DQ problems.* |
| 7 | *Improper selection of record granularity may lead to poor schema design and thereby affecting DQ.* |
| 8 | *Incomplete/Wrong identification of facts/dimensions, bridge tables or relationship tables or their individual relationships contribute to DQ problems.* |
| 9 | *Inability to support database schema refactoring cause data quality problems.* |

ISSN (Online): 1694-0784
ISSN (Print): 1694-0814

| | |
|---|---|
| *10* | *Lack of sufficient validation, and integrity rules in schema contribute to poor data quality.* |

### 2.3 Discussion on Classification

The paper has traced out the possible causes of data quality problems at every stage of data warehousing. Talking about classification of data quality issues at data sources stage, a set of almost 52 causes is framed which contribute towards data quality of the data warehouse if not taken care of may lead to poor data quality. This stage is considered to be more vulnerable to data quality problems as the data is culled from various heterogeneous environments. The most common type of problems that are manifested in literature of data quality are: lack of standardization of data, non standardization of formats, heterogeneity of data sources, Non-Compliance of data in data sources with the standards, missing data , inconsistent data across the sources, inadequate data quality testing on individual data source and many more depicted in Table1.

Data profiling is the technical analysis of data to describe its content, consistency and structure. Profiling can be divided into following categories [22]

*Pattern Analysis* – Expected patterns, pattern distribution, pattern frequency and drill down analysis

*Column Analysis* – Cardinality, null values, ranges, minimum/maximum values, frequency distribution and various statistics.

*Domain Analysis* – Expected or accepted data values and ranges

Table 2 formulated the classification of causes of data quality issues pertaining to the above classification of data profiling activities. A brief set of 19 possible causes of data quality problems are identified at this stage of data profiling.

ETL and data staging is considered to be more crucial stage of data warehousing process where most of the data cleansing and scrubbing of data is done. There can be myriad of reasons at this stage which can contribute to the data quality problems. A common mistake is to write custom programs to perform the extraction, transformation, and load functions. Writing an ETL program by hand may seem to be a viable option because the program does not appear to be too complex and programmers are available [23]. However, there are serious problems with hand-coded ETL programs. It give rise to problems like metadata is not generated automatically by hand-generated ETL programs, To keep up with the high volume of changes initiated by end users, hand-written ETL programs have to be constantly modified and in many cases rewritten, and much more problems can be there which may contribute to the data quality problems. Some of the problems are related to the capabilities of off self ETL tools use for the purpose of data warehousing and some are the traditional problems related to this phase. A brief set of 36 such issues have been identified at this stage and are mentioned in table 3.

Data Warehouse schema design and modeling is also considered to be vulnerable for contributing towards data quality problems. A flawed schema impacts negatively on information quality. Taking care of some of the factors related to data quality while designing the schema for warehouse would really help in achieving quality data in it. Table 4 depicts a set of 10 causes which if not taken care of while you design schema for warehouse would really deteriorate the data quality.

### 2.4 Conclusions

Today, to the best of our knowledge, no comprehensive & descriptive classification of causes of data quality problems exists. In this paper we have attempted to collect all possible causes of data quality problems that may exist at all the phases of data warehouse. Our objective was to put forth such a descriptive classification which covers all the phases of data warehousing which can impact the data quality. The motivation of the research was to integrate all the sayings of different researches which were focused on individual phases of data warehouse. Such as lot of literature is available on dirty data taxonomies, even some researchers have attempted to provide brief set of issues of data quality problems as well. But none of the research has attempted to think on near possible set of causes of data quality problems at all the phases at one attempt. Our classification of causes will really help the data warehouse practitioners, implementers and researchers for taking care of these issues before moving ahead with each phase of data warehousing. It would also be helpful for the vendors and those who are involved in development of data quality tools so as to incorporate changes in their tools to overcome the problems highlighted in classifications

### 2.5 Future work

Each item of the classification shown in table 1,2,3 and 4, will be converted into a item of the research instrument such as questionnaire and will be empirically tested by collecting views about these items from the data warehouse practitioners, appropriately.

**Acknowledgements**

Much of the information in this paper was derived from extended conversations with Subject Matter Experts

ISSN (Online): 1694-0784
ISSN (Print): 1694-0814

(SME), data warehouse practitioners, Data Quality Stewards of various IT companies of India and a through literature review of data quality issues. The authors gratefully acknowledge the time investments in this project that were generously provided by Mr.Saurabh Chopra, SME, AMDOCS, Pune India, Ms. Ravneet Bhatia and Ms. Rosy Choudhary Software Engineer (Data Warehousing Wing). ebusinessware India Pvt. Ltd. , Mr. Gursimran Singh Sr.Software Engineer (DBA) IBM India Pvt Ltd, Gurgaon, India. Special Thanks to people from abroad Mr. Girish Butaney and Mr. Nirlap Vora, Mr. Kerri Patel of SAS, Mr. Won Kim, Mr. Thilini Ariyachandra Assistant Professor of Management Information Systems Williams College of Business, Xavier University, Professor Yair Wand Head at the University of British Columbia, Faculty of Commerce and Business Administration, in Vancouver. *Jean-Pierre Dijcks* Principal Product Manager. Oracle Warehouse Builder. Oracle Corporation.

**References**

[1] Channah F. Naiman, Aris M. Ouksel (1995) "A Classification of Semantic Conflicts in Heterogeneous Database Systems", Journal of Organizational Computing, Vol. 5, 1995

[2] John Hess (1998), "Dealing With Missing Values In The Data Warehouse" A Report of Stonebridge Technologies, Inc (1998).

[3] Jaideep Srivastava, Ping-Yao Chen (1999) "Warehouse Creation-A Potential Roadblock to Data Warehousing", IEEE Transactions on Knowledge and Data Engineering January/February 1999 (Vol. 11, No. 1) pp. 118-126

[4] Amit Rudra and Emilie Yeo (1999) "Key Issues in Achieving Data Quality and Consistency in Data Warehousing among Large Organizations in Australia", Proceedings of the 32nd Hawaii International Conference on System Sciences – 1999

*[5]* Jesús Bisbal et all (1999) "Legacy Information Systems: Issues and Directions", *IEEE Software September/ October 1999*

*[6]* Scott W. Ambler (2001) "Challenges with legacy data: Knowing your data enemy is the first step in overcoming it", Practice Leader, Agile Development, Rational Methods Group, IBM, 01 Jul 2001.

[7] Scott W. Ambler "The Joy of Legacy Data" available at: http://www.agiledata.org/essays/legacyDatabases.html

[8] Wayne Eckerson, *"Data Profiling: A Tool Worth Buying (Really!)" ,Information Management Magazine, June 1, 2004 available http://wwwInformation-management.com/issues/20040601/1003990-1.html*

[9] "Boosting Data Quality for Business Success", INFORMATICA White Paper, 2006

[10] Federico Zoufaly *"Issues and Challenges Facing Legacy Systems ", available at http://www.developer.com/mgmt/article.php/11085_1492531*

[11] Won Kim et al (2002)- *"A Taxonomy of Dirty Data " Kluwer Academic Publishers 2002*

[12] Erhard Rahm & Hong Hai Do (2003) *"Data Cleaning: Problems and Current Approaches "*, available at: http://homepages.inf.ed.ac.uk/wenfei/tdd/reading/cleaning.pdf.

[13] HarteHanks *"Solving the source data problems with automated data profiling" Trillium Software, available at www.trilliumsoftware.com*

[14] Wayne Eckerson & Colin White (2003) *" Evaluating ETL and Data Integration Platforms" TDWI report series*

[15] Wayne W. E. (2004) *"Data Quality and the Bottom Line: Achieving Business Success through a Commitment to High Quality Data "*, The Data warehouse Institute (TDWI) report , available at www.dw-institute.com .

[16] Mike(2009), " "The problem of dirty data" available at http://www.articlesbase.com/databases-articles/the-problem-of-dirty-data-1111299.html

[17] The Standish Group (1999), "Migrate Headaches," available at *www.it-cortex.com/start_failure_rate.htm*

[18] Ralaph Kimball, The Data Warehouse ETL Toolkit, Wiley India (P) Ltd (2004),

[19] Tech Notes (2008), Why Data Warehouse Projects Fail: Using Schema Examination Tools to Ensure Information Quality, Schema Compliance, and Project Success. Embarcadero Technologies. Available at www.embarcadero.com.

[20] Markus Helfert, Gregor Zellner, Carlos Sousa, "Data Quality Problems and Proactive Data Quality Management in Data-Warehouse-Systems",

[21] David Loshin , "The Data Quality Business Case: Projecting Return on Investment", Informatica White paper. Available at :

ISSN (Online): 1694-0784
ISSN (Print): 1694-0814

http://www.melissadata.com/enews/articles/1007/2.htm

[22] Amol Shrivastav,Mohit Bhaduria, Harsha Rajwanshi (2008), " Data Warehouse and Quality Issues", available at http://www.scribd.com/doc/9986531/Data-Warehouse-and-Quality-Issues

[23] Ahimanikya Satapathy, "Building an ETL Tool", Sun Microsystems, Available at : http://wiki.open-esb.java.net/attach/ETLSE/ETLIntroduction.pdf

[24] Arkedy Maydanxhik (2007), " Causes of Data Quality Problems", Data Quality Assessment, Techniques Publications LLC. Available at http://media.techtarget.com/searchDataManagement/downloads/Data_Quality_Assessment_-_Chapter_1.pdf

**Ranjit Singh** is a research fellow with university college of Engineering & Technology (UCoE), Punjabi University Patiala, Punjab (INDIA). He is currently working on his PhD, "Development of Data Quality Assurance Model (DQAM) for Data Warehouse and its Impact on Business Intelligence". He has been awarded his Master of Computer Applications (M.C.A.) degree in 2002 from G.N.D.U, Amritsar, Punjab, India, and Master of Philosophy (M.Phil.) in Computer Science in 2006 from M.K.U. Madurai, Tamil Nadu, India. Presently author is working as senior lecturer in department of computer applications at Apeejay Institute of Management, Jalandhar, Punjab, India. He has authored 4 books in the area of computer applications and more than 6 research papers in national level journals and presented more than 10 papers in conference seminars at national level.

**Dr. Kawaljeet Singh** is working as Director, University Computer Center (UCC) at Punjabi University Patiala, Punjab, Inida. He has earned his Master of Computer Applications (MCA) in year 1988 and Doctorate of Philosophy (Ph.D.) in year 2001 in computer science from Thaper Institute of Engineering & Technology (T.I.E.T.), Deemed University, Patiala Punjab, India. He is having near about two decades of professional, administrative and academic career. He has served for more than a decade at various positions in Guru Nanak Dev University (GNDU) Amritsar, Punjab, India such as Reader & Head in Department of Computer Science and Engineering (DCSCE) , Amritsar campus, Professor and Head of the DCSCE & Electronics in Regional Campus of GNDU Jalandhar. He has also headed many responsible positions such as Dean & Chairman of Research Degree of the Committee of Faculty of Engineering & Technology of G.N.D.U. He has also co-authored 3 text books and published more than 32 research papers, abstracts, key notes and articles in national and international journals, conferences, seminars and workshops.

# IJCSI CALL FOR PAPERS NOVEMBER 2010 ISSUE

**Volume 7, Issue 6**

The topics suggested by this issue can be discussed in term of concepts, surveys, state of the art, research, standards, implementations, running experiments, applications, and industrial case studies. Authors are invited to submit complete unpublished papers, which are not under review in any other conference or journal in the following, but not limited to, topic areas. See authors guide for manuscript preparation and submission guidelines.

**Accepted papers will be published online and authors will be provided with printed copies and indexed by Google Scholar, Cornell's University Library, ScientificCommons, CiteSeerX, Bielefeld Academic Search Engine (BASE), SCIRUS and more.**

**Deadline: 30th September 2010**
**Notification: 31st October 2010**
**Revision: 10th November 2010**
**Online Publication: 30th November 2010**

- Evolutionary computation
- Industrial systems
- Evolutionary computation
- Autonomic and autonomous systems
- Bio-technologies
- Knowledge data systems
- Mobile and distance education
- Intelligent techniques, logics, and systems
- Knowledge processing
- Information technologies
- Internet and web technologies
- Digital information processing
- Cognitive science and knowledge agent-based systems
- Mobility and multimedia systems
- Systems performance
- Networking and telecommunications
- Software development and deployment
- Knowledge virtualization
- Systems and networks on the chip
- Context-aware systems
- Networking technologies
- Security in network, systems, and applications
- Knowledge for global defense
- Information Systems [IS]
- IPv6 Today - Technology and deployment
- Modeling
- Optimization
- Complexity
- Natural Language Processing
- Speech Synthesis
- Data Mining

**For more topics, please see http://www.ijcsi.org/call-for-papers.php**

All submitted papers will be judged based on their quality by the technical committee and reviewers. Papers that describe research and experimentation are encouraged.
All paper submissions will be handled electronically and detailed instructions on submission procedure are available on IJCSI website (www.IJCSI.org).

For more information, please visit the journal website (www.IJCSI.org)

www.IJCSI.org

# IJCSI

The International Journal of Computer Science Issues (IJCSI) is a refereed journal for scientific papers dealing with any area of computer science research. The purpose of establishing the scientific journal is the assistance in development of science, fast operative publication and storage of materials and results of scientific researches and representation of the scientific conception of the society.

It also provides a venue for researchers, students and professionals to submit on-going research and developments in these areas. Authors are encouraged to contribute to the journal by submitting articles that illustrate new research results, projects, surveying works and industrial experiences that describe significant advances in field of computer science.

**Indexing of IJCSI:**

1. Google Scholar
2. Bielefeld Academic Search Engine (BASE)
3. CiteSeerX
4. SCIRUS
5. Docstoc
6. Scribd
7. Cornell's University Library
8. SciRate
9. ScientificCommons


**www.IJCSI.org**