ICIQ 2010

http://iciq2010.org

November 12-14, 2010

Little Rock, Arkansas

The 15th International Conference on Information Quality

ICIQ 2010

http://iciq2010.org/

November 12-14, 2010

Hosted by

The Donaghey College of Engineering and Information Technology

University of Arkansas at Little Rock

Table of Contents

Welcome Message from the ICIQ 2010 Conference & Program Chairs

Dear Conference Participants,

Welcome to the 15th International Conference on Information Quality (ICIQ 2010). Each year ICIQ provides a forum for academics and practitioners from around the world to share their research findings and knowledge in order to expand the Information Quality (IQ) discipline. This year's conference offers sessions on such topics as IQ frameworks, tools, assessment methods, data architectures, web applications, cases from the financial and healthcare sectors, as well as technologies impacting IQ such as high performance computing and data sensors. ICIQ 2010 also features three keynote talks, poster presentations, sponsor exhibits, and ample opportunities for informal discussion and networking.

This is the first time that the ICIQ conference is being hosted by the University of Arkansas at Little Rock. From 1996 to 2008, the ICIQ conferences were held annually at MIT in Cambridge, Massachusetts. In 2009, the ICIQ conference went abroad when it was hosted by the Hasso Plattner Institute (HPI) in Potsdam, Germany. This year we are proud to showcase Arkansas' capital city, Little Rock.

Our downtown bustling River Market District features restaurants, shops, bars, museums and hotels. Two focal points in the area are the Clinton Presidential Center & Park and the neighboring world headquarters of Heifer International. A recent addition is the Central Arkansas Nature Center. The arts flourish in Little Rock with the Arkansas Arts Center, housing one of the most-acclaimed collections of works on paper in the country. In addition to the Clinton Presidential Center, some of the other historic sites include Little Rock Central High School National Historic Site, Historic Arkansas Museum, Old State House Museum, Mount Holly and MacArthur Museum of Military History. Little Rock's position in the center of the state also puts it on such historic trails as the Butterfield Overland Trail, the Southwest Trail, the Trail of Tears, and the Little Rock Campaign of the Civil War.

We hope you enjoy this year's ICIQ 2010 conference and your visit to Little Rock!

John Talburt and Elizabeth Pierce

The UALR-IQ Graduate Program at a Glance

The Information Quality (IQ) Graduate Program at the University of Arkansas at Little Rock (UALR) prepares students to pursue a variety of IQ careers such as Information Quality Manager, Information Quality Analyst, Data Management Consultant, or Data Architect. This program also prepares students to pursue doctoral-level graduate studies in preparation for information quality research and instructional roles. Established in 2006 by UALR's Donaghey College of Engineering and Information Technology (EIT) in collaboration with the Massachusetts Institute of Technology Information Quality (MIT IQ) Program, the UALR Information Quality Graduate Program is dedicated to promoting state of the art curriculum in information quality education, contributing new ideas to the information quality knowledge base, and establishing information quality partnerships with the community, government, and industry.

The UALR Information Quality Graduate Program is designed to meet the growing demands by government, industry, and non-profit organizations for qualified professionals with graduate-level degrees who can understand and apply:

- Concepts, principles, tools, and models essential in defining, measuring, analyzing, and improving the quality of data as judged by its fitness for use in a particular application
- Development of information quality strategies, policies, and programs to support an organization's operational, tactical, and strategic needs
- Interrelationships between information quality and other key information issues such as data privacy and protection, enterprise architecture, data mining, and data integration processes including identity resolution and customer relationship management
- Information Science theories and practices in the areas of database systems, systems analysis, and information visualization
- A learning environment that promotes critical thinking, communication skills, and project management

Students can choose from three different graduate degree options:

1. a 12 credit hours Graduate Certificate in Information Quality,
2. a 33 credit hours Master of Science in Information Quality, or
3. a 72 credit hours PhD in Integrated Computing with an Emphasis in Information Quality.

To learn more about the Information Quality Graduate Program at UALR, please check out the website: http://ualr.edu/informationquality/.

CONFERENCE SCHEDULE SUMMARY

<table>
<tr><td colspan="5">Friday, November 12 (Downtown Little Rock)</td></tr>
<tr><td>8:00-5:00 pm</td><td colspan="4">Reserved for IAIDQ Meeting – by invitation only.</td></tr>
<tr><td>5:00-6:30 pm</td><td colspan="4">Preconference Reception, Acxiom River Market Tower located at 601 East 3rd Street, Little Rock, AR 72201</td></tr>
<tr><td colspan="5">Saturday, November 13 (EIT Building on the UALR Campus, 2801 South University Avenue, Little Rock, AR 72204)</td></tr>
<tr><td>Bus Transportation</td><td colspan="4">Departs Peabody Hotel for UALR Campus at 7:40 am and 8:20 am</td></tr>
<tr><td>8:00-9:00 am</td><td colspan="3">Registration, Morning Refreshments, EIT Lobby</td><td rowspan="6">MSIQ Project Presentations (EIT 217)</td></tr>
<tr><td>9:00-10:30 am ICIQ 2010 Opening Ceremonies in EIT Auditorium</td><td colspan="3">• 9:00 – 9:25 Conference Welcome and Recognitions
• 9:25 – 9:30 Best ICIQ Paper Award
• 9:30 – 9:35 Keynote Introduction for Cita M. Furlani
• 9:35 – 10:20 Keynote Address: “Information and Software Quality, Integrity, and Usability from the Viewpoint of the National Institute of Standards and Technology”
• 10:20-10:30 Questions/Answers</td></tr>
<tr><td>10:30-10:45 am</td><td colspan="3">Break</td></tr>
<tr><td></td><td>EIT 224</td><td>EIT 218</td><td>EIT 220</td></tr>
<tr><td>10:45-12:15 pm Parallel Sessions 1</td><td>Finance Session, Chair: David Becker
• A Novel Approach to the Evaluation and Improvement of Data Quality in the Financial Sector
• Enabling Transparency & Trust in Financial Data Through Semantic Data Quality Rating System
• Using a Metro-Ethernet Project to Address Information Quality Challenges with Technology</td><td>Assessment Session, Chair: Jing Gao
• A Decision Rule Method for Data Quality Assessment
• A Hybrid Approach to Assessing Data Quality
• Subjective Evaluation of Perception of Accuracy in Visualization of Data</td><td>Healthcare Session, Chair: Bruce Davidson
• Assessing Information Quality Deficiencies in Emergency Medical Service Performance
• Assuring Information Quality of Electronic Health Records in eHealth Platform
• Tracing Infectious Diseases in South America: An Information Quality Challenge</td></tr>
<tr><td>12:15-1:45 pm Luncheon in EIT Lobby/Auditorium</td><td colspan="3">Speaker Introduction for Rod Ford, President & CEO of CognitiveDATA, Inc.
Luncheon Address: “CognitiveDATA – The Story of a Data Quality Start-Up”</td></tr>
</table>

CONFERENCE SCHEDULE SUMMARY

Saturday, November 13 (UALR Campus, EIT Building)				
	EIT 224	**EIT 218**	**EIT 220**	
1:45-3:15 pm Parallel Sessions 2	Invited Speaker I Session, Danette McGilvray • *The Business Value from Data Quality*	Architecture Session, Chair: Andy Koronios • *Enterprise Master Data Architecture: Design Decisions and Options* • *Corporate Data Quality Management in Context* • *Master Data Management: A Proof of Concept*	Sensor Data Session, Chair: G. Shankaranarayanan • *Quality of Information in Wireless Sensor Networks: A Survey* • *Improving Environmental Sensor Data Quality Using a Categorization of Data Properties* • *Dempster-Shafer Based Information Quality Processing in Smart Environments*	**Sponsor Exhibits (EIT Lobby)**
3:15-3:30 pm	**Break**			
3:30-4:30 pm Parallel Sessions 3	Web Session, Chair: Robert Pokorny • *Information Quality Challenges in Social Media* • *A Framework for Developing Better Instruments for Web-Based Health Information Quality*	IQ Models I Session, Chair: Beverly Kahn • *Information Quality for Operations: Framework and Model* • *Using Information Capacity to Assess Integrated Schema*	Invited Speaker II Session, Christian Walenta • *The Job of the Information / Data Quality Professional*	**ICIQ Poster Sessions (EIT Lobby)**
4:30-6:00 pm	**Opening Day Reception at the Bailey Center**			
Bus Transportation	**Departs UALR Campus for Peabody Hotel at 5:20 pm and 6:00 pm**			

CONFERENCE SCHEDULE SUMMARY

Sunday, November 14 (Peabody Hotel, 3 Statehouse Plaza, Little Rock, AR 72201)			
8:00-9:30 am Breakfast in Peabody Hotel's Riverview Room	Speaker Introduction for Sharif Youssef and Rick McConville, Acxiom Breakfast Address: "How Consumer Trends are Shifting the Data Quality Landscape"		
	Chico	**Grampas**	**Ouachita**
9:30-10:30 am Parallel Sessions 4	Data Migration Session, Chair: C. Lwanga Yonke • *Data Quality as a Key Success Factor for Migration Projects* • *Migrating and Cleaning Data Using Excel: A Case Study*	IQ Models II Session, Chair: Farhad Moeeni • *A Discussion of Methods for the Detection of Errors in a Power Law Distribution* • *A General Ranking Strategy for Data Accuracy Management*	IQ Tools Development Session, Chair: Andrea Maurino • *Snap-On Data Quality Enhancement and Verification Tool (DEVA) for Asset Management* • *A New Cycle of Improvement for Information Quality Services*
10:30-10:45 am	**Break**		
10:45-11:45 am Parallel Sessions 5	IQ Fundamentals Session, Chair: Ningning Wu • *Framing Data Quality Research: A Semantic Analysis Approach* • *Analysis of Information Quality Criteria in a Crisis Situation as a Characteristic of Complex Situations*	IQ Models III Session, Chair: Mariofanna Milanova • *The Effect of Data Quality on Data Mining – Improving Prediction Accuracy by Generic Data Cleansing* • *Managing Accuracy of Project Data in a Distributed Project Setting*	High Performance Computing Session, Chair: Valerie Sessions • *Towards a High Performance Merge Solution for Large Scale Datasets* • *Towards a Data Intensive Approach to Named Entity Recognition*
12:00 pm	**Official End of ICIQ 2010 Conference Program**		
12:00-2:00 pm	**Optional Brunch and Tour of the Clinton Presidential Library**		

ICIQ 2010 Organization

Conference Chairs
John R. Talburt, University of Arkansas at Little Rock (jrtalburt@ualr.edu)
Andy Koronios, University of South Australia, Adelaide (Andy.Koronios@unisa.edu.au)
Ying Su, Institute of Scientific and Technical Information of China, Beijing (suy.istic@gmail.com)

Program Chairs
Elizabeth Pierce, University of Arkansas at Little Rock (expierce@ualr.edu)
Ismael Caballero Muñoz-Reja, Universidad de Castilla-La Mancha (Ismael.Caballero@uclm.es)
Anne Marie Smith, EWSolutions/North Centeral University (asmith@manieri.com)
Diane Strong, Worcester Polytechnic Institute (dstrong@wpi.edu)

Program Committee
Nitin Agarwal, University of Arkansas at Little Rock
Carlo Batini, University of Milan
David Becker, MITRE Corporation
Laure Berti-Equille, Université de Rennes
Michael Boggs, Analytix
Paul Bowen, Florida State University
Mikhaila Burgess, Cardiff University
Lan Cao, Old Dominion University
Cinzia Cappiello, Politecnico di Milano
InduShobha Chengalur-Smith, SUNY at Albany
Chia-Chu Chiang, University of Arkansas at Little Rock
Serhan Dagtas, University of Arkansas at Little Rock
Tamraparni Dasu, AT&T Labs - Research
Bruce Davidson, Cedars-Sinai Health System
Adir Even, Ben-Gurion University of the Negev
Craig Fisher, Marist College
Zbigniew J Gackowski, California State University Stanislaus
Jing Gao, University of South Australia
Michael Gertz, Ruprecht-Karls-University Heidelberg
Olaf Hartig, Humboldt University of Berlin
Markus Helfert, Dublin City University
Beverly Kahn, Suffolk University
Stephen Kennet, Australia Department of Defense - DSTO
Barbara Klein, University of Michigan at Dearborn
Akihisa Kodate, Tsuda College
Hiroshi Koga, Kansai University
Eitel Lauria, Marist College
Yang Lee, Northeastern University
Peggy.Leonowich-Graham, USMA
Philip Lesslar, IAIDQ

Jiuyong Li, University of South Australia
Bart Longenecker, University of South Alabama
Helina Melkas, Lappeenranta University of Technology
Michael Mielke, DB Mobility Logistics AG
Mariofanna Milanova, University of Arkansas at Little Rock
Paolo Missier, The University of Manchester
Farhad Moeeni, Arkansas State University
Paulo Jorge Oliveira, Politécnico do Porto
Barbara Pernici, Politecnico di Milano
Leo Pipino, University of Massachusetts Lowell
Robert Pokorny, XSB, Inc.
Srini Ramaswamy, University of Arkansas at Little Rock
Mary Ann Robbert, Bentley College
Grant Robinson, New South Wales Office of Water
Shazia Sadiq, The University of Queensland
Kai-Uwe Sattler, Ilmenau University of Technology
Monica Scannapieco, Italian National Institute of Statistics (Istat)
Scott Schumacher, IBM
Thomas A. J. Schweiger, Acxiom
Laura Sebastion-Coleman, Ingenix
Yasuki Sekiguchi, Hokkaido University
Valerie Sessions, Charleston Southern University
Ganesan Shankaranarayanan, Babson College
John (Skip) Slone, Lockheed Martin Corp.
Terry Talley, Southwest Power Pool
Giri Kumar Tayi, SUNY at Albany
Mihail E. Tudoreanu, University of Arkansas at Little Rock
Niels Weigel, SAP AG
Rolf Wigand, University of Arkansas at Little Rock
Harris Wu, Old Dominion University
Ningning Wu, University of Arkansas at Little Rock
C. Lwanga Yonke, Aera Energy LLC
Harry Zhu, Old Dominion University

Local Arrangements Chair
Kelley Bass, University of Arkansas at Little Rock (wkbass@ualr.edu)

Planning and Advisory Committee
Brenda Barnhill, Chair, University of Arkansas at Little Rock (bkbarnhill@ualr.edu)
John R. Talburt, University of Arkansas at Little Rock (jrtalburt@ualr.edu)
Elizabeth Pierce, University of Arkansas at Little Rock (expierce@ualr.edu)
Kelley Bass, University of Arkansas at Little Rock (wkbass@ualr.edu)
Rick McGraw, BlackOak Partners (mcgraw@blackoakgroup.com)
Larry Mougeot, BlackOak Partners (lmouge@blackoakgroup.com)
Gregg Webster, University of Arkansas at Little Rock (glwebster@ualr.edu)
Thomas Wallace, University of Arkansas at Little Rock (tswallace@ualr.edu)

Keynote Speakers

Cita M. Furlani
Director of the Information Technology Laboratory (ITL)
National Institute of Standards and Technology (NIST)

ICIQ 2010 Opening Keynote Speaker
Saturday, November 13, 2010

"Information and Software Quality, Integrity and Usability from the Viewpoint of the National Institute of Standards and Technology"

Bio: Cita M. Furlani is Director of the Information Technology Laboratory (ITL). ITL is one of six research Laboratories within the National Institute of Standards and Technology (NIST) with an annual budget of $120 million, 367 employees, and about 160 guest researchers from industry, universities, and foreign laboratories.

Furlani oversees a research program designed to promote U.S. innovation and industrial competitiveness by developing and disseminating standards, measurements, and testing for interoperability, security, usability, and reliability of information systems, including cybersecurity standards and guidelines for Federal agencies and U.S. industry, supporting these and measurement science at NIST through fundamental and applied research in computer science, mathematics, and statistics. Through its efforts, ITL seeks to enhance productivity and public safety, facilitate trade, and improve the quality of life.

Within NIST's traditional role as the overseer of the National Measurement System, ITL is addressing the hard problems in IT Measurement Research. ITL's research results in metrics, tests, and tools for a wide range of subjects such as complex systems, pervasive information technologies, and virtual measurements, as well as issues of information and software quality, integrity, and usability.

ITL has been charged with leading the nation in utilizing existing and emerging IT to meet national priorities that reflect the broad-based social, economic, and political values and goals of the country. Under the Federal Information Security Management Act, ITL is charged with

developing cybersecurity standards, guidelines, and associated methods and techniques. Under other legislation, such as the USA PATRIOT Act, the Help America Vote Act, and the American Recovery and Reinvestment Act, ITL is addressing the major challenges faced by the nation in the areas of homeland security, electronic voting, and health information technology.

Furlani has served as the Acting Director of the NIST Advanced Technology Program and as Chief Information Officer for NIST. She previously served as director of the National Coordination Office for Networking and Information Technology Research and Development. This office, reporting to the White House through the Office of Science and Technology Policy and the National Science and Technology Council, coordinates the planning, budget, and assessment activities for the Networking and Information Technology Research and Development Program.

She has been awarded the Department of Commerce Silver and Bronze Medal Awards.

Keynote Speakers

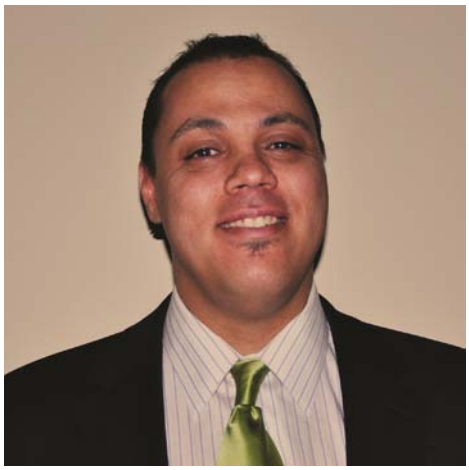
Rod Ford
President & CEO
CognitiveDATA, Inc.

Luncheon Speaker
Saturday, November 13, 2010

"CognitiveDATA – The Story of a Data Quality Start-Up"

Bio: Rod Ford is an innovator and subject matter expert in the field of data quality and marketing data accuracy. As the Founder & CEO of CognitiveDATA, his technology innovation and executive leadership pushed the company into a widely recognized position of thought leadership for data quality accuracy in the direct marketing services industry. Rod Ford is a mechanical engineer by degree. CognitiveDATA is the 4th technology company he has launched in the past 22 years and the company recently celebrated its ninth year anniversary by being named for the third consecutive year by Inc. Magazine as one of the fastest growing private companies in the U.S. The company's data accuracy technology is utilized by over 500 direct marketing companies including industry captains AT&T, Geico, DirecTV, Gap, and Hartford. In the spring of last year Rod sold the company to Merkle, Inc., and today CognitiveDATA operates as a wholly owned subsidiary of the parent company, with over 100 employees and revenues in excess of $20 million. His blog can be viewed at http://blog.cognitivedata.net.

Keynote Speakers



Sharif Youssef
Enterprise Product Strategist
Acxiom

Breakfast Speaker
Sunday, November 14, 2010

"How Consumer Trends are Shifting the Data Quality Landscape"

Bio: Sharif Youssef is an Enterprise Product Strategist with expertise in identity management and multi-channel marketing for Acxiom. He leads strategy for the innovation and evolution of several existing Acxiom products, including multi-channel marketing, identity management and risk mitigation products. He is a key consultant on multi-layer marketing and risk mitigation strategy, needed in today's online environment.

As a thought leader in developing new ways to optimize ROI for enterprise product strategy development, reduce client fraud and create multi-channel marketing solutions, Sharif has a combination of skills not commonly found in the data space. He possesses a clear vision of risk mitigation and marketing trends, plus the ability to render them operational.

Sharif has been in the technology industry for over 14 years working with some of the largest data companies such as First Data, Experian, TransUnion, and IBM, responsible for the creation and management of ground breaking solutions ranging from online tax payment systems deployed by the Federal Government, to eCommerce platforms, multi-channel marketing and identity/risk mitigation solutions. His accomplishments include eFile, System Architect for COLINK and the development of many proprietary solutions for leading Internet companies.

Sharif lives in Denver, CO with several hardy plants that can survive his travel schedule.

Keynote Speakers

Rick McConville
Product Manager for PanOptic-X
Suite of Data Quality Products
Acxiom

Breakfast Speaker
Sunday, November 14, 2010

"How Consumer Trends are Shifting the Data Quality Landscape"

Bio: Rick McConville is the Product Manager for Acxiom's PanOptic-X suite of data quality products, for which he drives the financial and market success. He spearheads the development of requirements and merchandizing strategies, as well as executing competitive, financial, market and internal analysis on core product capabilities for strategic sales initiatives.

Rick's expertise in data quality comes from 15 years of Postal Industry and Data Quality experience. Prior to Acxiom, he managed Data Licensing at the United States Postal Service, National Customer Support Center (NCSC) in Memphis TN, where he led Address Management projects ranging from CASS and MASS to NCOALink licensing. In his 13 year postal career he was awarded over 20 times for outstanding leadership.

Rick represents Acxiom on the Board of Directors for the Association for Postal Commerce (PostCom). He is often called on to present to numerous industry organizations on a variety of topics related to postal and data quality.

Rick attended the University of Memphis and lives in Memphis, TN with his lovely wife, Amy, and their two miniature dachshunds.

Research Presentations

Parallel Session 1-A Saturday, November 13, 2010 10:45 am to 12:15 pm	Finance Session Chair: David Becker EIT 224

A Novel Approach to the Evaluation and Improvement of Data Quality in the Financial Sector
Karel Dejaeger, Bart Hamersb, Jonas Poelmansa, and Bart Baesensa

Abstract: Until recently, a universal approach for analyzing the quality of data generated by business processes has been missing. In this paper, a structured approach to data quality management is presented and applied to the credit rating process of a company in the financial sector. Starting from a comprehensive data quality definition, a structured questionnaire is composed. It is used for guided interviews to distill the key elements in the credit rating process from a data quality perspective. Business Process Modeling Notation (BPMN) is used to visualize the process, to identify where data elements enter the process, and to trace the various data outputs of the process. The visual process representation allows one to identify error prone elements in the process and thus enabling a more focused data quality management. It was found that data quality issues appear manifold in larger data driven organizations in the financial sector. The presented methodology provides a cross-departmental overview of the data flow and manipulations from source to end usage.

Enabling Transparency & Trust in Financial Data Through Semantic DQ Rating System
Ashu Bhatnagar

Abstract: Both, the government regulators and the investors in global capital markets rely heavily on reported financial data. This financial data enters the information supply chain as raw data in a variety of formats and reported by the companies as quarterly and annual filing data. Similarly, stock exchanges provide high volume, real-time market data, including price and trading volumes. Error-free, high quality data from these sources is critical for compliance and investment management decisions. At the same time, an intense focus on outlier data is of great interest to both of these groups as it may represent a potential compliance issue or a potential arbitrage opportunity for investment decision, known as an 'Alpha' opportunity. Currently, this screening for errors, for most part, is a manual process, and given the vast amount of data, prone to human errors. Extremely rare, but nonetheless real, so called fat-finger errors continue to cast a large shadow of doubt on overall data integrity and undermine trust in the system. This score, now metadata, is linked with other contextual and semantic metadata. Aggregating cell level quality score then enables visual display of potential outliers as hot spots on a heat map. A drill-down using linked data increases transparency in data, all the way to the original source enabling trust in the overall data set.

Using a Metro-Ethernet Project to Address IQ Challenges with Technology
James McGinnis and Bruce Upton

Abstract: Metro Ethernet is defined as a "computer network that covers a metropolitan area and that is based on the Ethernet standard. It is commonly used as a metropolitan access network to connect subscribers and businesses to a larger service network or the Internet. Businesses can also use Metro Ethernet to connect branch offices to their Intranet". This paper details the recent move from Multiprotocol Layer Switching (MPLS) leased data circuits to Metro Ethernet by a financial institution. While the product implementation discussed below is not an entirely new approach, the application of the technology to address and measure specific information quality dimensions should be useful to other researchers and other similar institutions seeking to measure information quality as company product. The authors have documented some of the project actions and results as they relate to IQ. The institution's overall goal was to use this goal to improve IQ.

Research Presentations

Parallel Session 1-B Saturday, November 13, 2010 10:45 am to 12:15 pm	Assessment Session Chair: Jing Gao EIT 218

A Decision Rule Method for Data Quality Assessment

Nawaf Alkharboush and Yuefeng Li

Abstract: The assessment of data quality is a key success factor for organizational performance. It supports managers and executives to clearly identify and reveal defective data in their information systems, and consequently minimizes and eliminates the risks associated with decisions based on poor data. Despite the importance of data quality assessment, limited research has been conducted on providing an objective data quality assessment. Researchers and practitioners usually rely on an error ratio metric to calculate abnormal data. However, this approach is insufficient in terms of providing a complete quality assessment since errors can be randomly and systematically distributed across databases. This study will introduce a decision rule method for providing a comprehensive quality assessment, which captures and allocates quality change at the early stage in organizational information systems. A decision rule can also be extended to answer important questions such as the randomness degree and the probability distribution of errors. These advantages will significantly reduce the time and costs associated with performing quality assessment tasks. More importantly, the efficiency and effectiveness of the decision rule for assessing data quality enables management to make accurate decisions reflecting positively on organizational values.

A Hybrid Approach to Assessing Data Quality

Philip Woodall and Ajith Kumar Parlikad

Abstract: Various techniques have been proposed to enable organizations to initiate procedures to assess and ultimately to improve the quality of their data. The utility of these assessment techniques (ATs) has been demonstrated in different organizational contexts. However, while some of the ATs are geared towards specific application areas and are often not suitable in different applications, others are more general and therefore do not always meet specific requirements. To address this problem we propose the Hybrid Approach to assessing data quality, which can generate usable ATs for specific requirements using the activities of existing ATs. A literature review and bottom-up analysis of the existing data quality (DQ) ATs was used to identify the different activities proposed by each AT. Based on example requirements from an asset management organization, the activities were combined using the Hybrid Approach in order to generate an AT which can be followed to assess an existing DQ problem. The Hybrid Approach demonstrates that it is possible to develop new ways of assessing DQ which leverage the best practices proposed by existing ATs by combining the activities dynamically.

Subjective Evaluation of Perception of Accuracy in Visualization of Data

Ahmed Abuhalimeh, M. Eduard Tudoreanu, and Erich A. Peterson

Abstract: This paper focuses on the human's perception of information quality and describes the results of a study on how accuracy is estimated for data shown through a visual representation. The subjective assessment of quality appears to be non-linear in relation to the actual degree of errors in the dataset. Users are sometimes unable to distinguish between datasets with different quality, and their ability to estimate is better for certain quality levels than for others. The study also shows that adding complementary information does not always help users to better assess the accuracy of the visualization, and thus of the data. The implication of these results is that, for subjective measures of quality, traditional statistical methods of assessing quality may need to be extended with additional methods to account for the non-linearity and the behavior of data integration.

Research Presentations

Parallel Session 1-C Saturday, November 13, 2010 10:45 am to 12:15 pm	Healthcare Session Chair: Bruce Davidson EIT 220

Assessing Information Quality Deficiencies in Emergency Medical Service Performance
Shuyan Xie and Markus Helfert

Abstract: There is an inevitable need for information sharing and coordinating among response departments during the occurrences of emergencies. Information quality in EMS performance has been the subject of very little research so far, especially on the influence of information deficiencies to EMS performance, despite the fact that the studied information quality in healthcare has been discussed. Due to the excessive amount of information and the dynamic change in the environment, the information decision process has become the backbone of EMS. Aim of this paper is to define dimensions for describing information quality deficiencies concerning the information flow across units from the communication center to dispatch center, to mobile rescue units, and to emergency department (ED). We present a literature review on information quality in emergency medical service (EMS) and identify and define the dimensions for describing information quality., Three case studies are presented in order to describe the coordination in emergency service performance through information exchange and sharing. The contribution lies in the defined and exemplified dimensions for describing information quality deficiencies as a step towards our future work on information integration model for EMS performance improvement. A gross list of eight dimensions were defined from literature and used in describing information quality deficiencies in EMS performance of three cases.

Assuring Information Quality of Electronic Health Records in eHealth Platform
Ying Su, Ling Yin, and Latif Al-Hakim.

Abstract: The core application of the eHealth is an electronic health record (EHR) system. One significant obstacle is the class of problems that arise due to variations in the quality of the information being shared. In this paper we outline a framework for assuring information quality (IQ) in the EHR context, using semiotics theory, semantic explanation of resources, and data couplings. Physicians can define the quality characteristics that are of importance in their particular domain by applying an IQ semiotics, which classifies and organizes these domain-specific quality characteristics within this quality assurance framework. Resource Description Framework (RDF) is used to explain data resources, with reference to IQ indicators defined in the semiotics. Data couplings - again defined in RDF - are used to represent mappings between data elements and the IQ semiotics. As a practical illustration of our approach, we present a case study from an open eHealth platform for the community-wide health information network of China.

Tracing Infectious Diseases in South America: An IQ Challenge
Maria Jose Espona

Abstract: Biological Weapon (BW) agents are pathogens present in the nature that were included in weapons programs, but at the end, they are just microorganisms that cause diseases around the globe. In this paper, we will show the aspects we consider are involved in the biological agents' epidemiology in South America and with special focus in Argentina, using open sources information, from an Information Quality (IQ) perspective. In this sense, and after study the data obtained, we found out several IQ problems, such as, several IQ dimensions are affected and cultural and organizational aspects deeply affect the disease reporting process. This is a research in progress; there is a long way ahead in this journey that tries to understand both the epidemiology of certain diseases and the disease reporting schema. Given the fact that in this research we deal with national and supranational organizations information, we propose some recommendations to them in order to improve their situation.

Research Presentations

Parallel Session 2-A **Saturday, November 13, 2010** **1:45 pm to 3:15 pm**	**Invited Speaker I Session** **Danette McGilvray** **EIT 224**

The Business Value from Data Quality

Danette McGilvray, President and Principal, Granite Falls Consulting Inc.

Abstract: Awareness of any data quality issue immediately leads to questions such as "What impact does information quality have on the business?" and "Why does data quality matter?" Historically it has been difficult to answer these and demonstrate the value of information quality. This presentation provides an overview of business impact techniques which are qualitative and quantitative methods for determining the effects of information quality on any organization. These approaches can be used in many situations - whether you are beginning an information quality program, implementing a data-quality focused project, including data quality in another project or methodology, or are an individual tasked with responsibilities in these areas. Techniques are applied based on need, time and resources available. A case study shows how a variety of the techniques were used to develop and present the "Business Value from Data Quality" at Sallie Mae, a Fortune 500 company and the United State's leading provider of saving, planning, and paying for education programs.

Research Presentations

Parallel Session 2-B Saturday, November 13, 2010 1:45 pm to 3:15 pm	Architecture Session Chair: Andy Koronios EIT 218

Enterprise Master Data Architecture: Design Decisions and Options
Boris Otto and Alexander Schmidt

Abstract: The enterprise-wide management of master data is a prerequisite for companies to meet strategic business requirements such as compliance to regulatory requirements, integrated customer management, and global business process integration. Among others, this demands systematic design of the enterprise master data architecture. The current state-of-the-art, however, does not provide sufficient guidance for practitioners as it does not specify concrete design decisions they have to make and to the design options of which they can choose with regard to the master data architecture. This paper aims at contributing to this gap. It reports on the findings of three case studies and uses morphological analysis to structure design decisions and options for the management of an enterprise master data architecture.

Corporate Data Quality Management in Context
Ana Lucas

Abstract: Presently, we are well aware that poor quality data is costing large amounts of money to corporations all over the world. Nevertheless, little research has been done about the way Organizations are dealing with data quality management and the strategies they are using. This work aims to find some answers to the following questions: which business drivers motivate the organizations to engage in a data quality management initiative?, how do they implement data quality management? and which objectives have been achieved, so far? Due to the kind of research questions involved, a decision was made to adopt the use of multiple exploratory case studies as research strategy. The case studies were developed in a telecommunications company (MyTelecom), a public bank (PublicBank) and in the central bank (CentralBank) of one European Union Country. The results show that the main drivers to data quality (DQ) initiatives were the reduction in non quality costs, risk management, mergers, and the improvement of the company's image among its customers, those aspects being in line with literature [7, 8, 20]. The commercial corporations (MyTelecom and PublicBank) began their DQ projects with customer data, this being in accordance with literature [18], while CentralBank, which mainly works with analytical systems, began with data source metadata characterization and reuse. None of the organizations uses a formal Data Quality methodology, but they are using tools for data profiling, standardization and cleaning. PublicBank and CentralBank are working towards a Corporate Data Policy, aligned with their Business Policy, which is not the case of MyTelecom. The findings enabled us to prepare a first draft of a "Data Governance strategic impact grid", adapted from Nolan& MacFarlan IT Governance strategic impact grid [17], this framework needing further empirical support.

Master Data Management: A Proof of Concept
Helena Galhardas, Luis Torres, and João Damásio

Abstract: Nowadays, organizations deal with multiple data sources. Therefore, they must cope with data quality problems such as inconsistent and inaccurate data, as well as with data integration and synchronization. In this paper, we present an implementation of an MDM architecture using the open source tool Mural. Furthermore, we report preliminary experimental results that demonstrate the advantages of an MDM system in a real world context.

Parallel Session 2-C Saturday, November 13, 2010 1:45 pm to 3:15 pm	Sensor Data Session Chair: G. Shankaranarayanan EIT 220

Quality of Information in Wireless Sensor Networks: A Survey
Vinay Sachidananda, Abdelmajid Khelil, Neeraj Suri.

Abstract: In Wireless Sensor Networks (WSNs) the operating conditions and/or user requirements are often desired to be evolvable, whether driven by changes of the monitored parameters or WSN properties of configuration, structure, communication capacities, node density, and energy among many others. While considering evolvability, delivering the required information with the specified quality (accuracy, timeliness, reliability etc) defined by the user constitutes a key objective of WSNs. Most existing research efforts handle fluctuations of operation conditions in order to deliver information with the highest possible specified quality. In this paper, we take these aspects into consideration and survey existing work on Quality of Information (QoI). As a contribution, we categorize WSN information into a set of abstract classes for generality across varied application types. Our survey shows that currently QoI is usually addressed in isolation by focusing on discrete data processing operations/building blocks such as raw data collection, in-network processing (compression, aggregation), information transport and sink operations for decision making. This survey comprehensively explains the different views of QoI on attributes, metrics and WSN functional operations mapped with existing approaches. The survey also forms the basis for specifying needed QoI research issues.

Improving Environment Sensor Data Quality Using a Categorization of Data Properties
Irbis Gallegos, Ann Gates, and Craig Tweedie

Abstract: To conduct research on the causes of global environmental changes, environmental scientists have been using advanced technologies, such as wireless sensor networks and robotic trams equipped with sensors, to collect spectral readings, ground temperature, ground moisture, wind velocity, light spectrum, and other data. Indeed, the amount of data being collected is rapidly increasing, and the ability to evaluate promptly the accuracy of the data and the correct operation of the instrumentation being used to collect the data is critical in order to not lose valuable time and information. To address these issues, an approach based on software-engineering techniques is being developed to support the scientist's ability to specify data properties, through guidance using property classifications, which can then be used for near real-time monitoring of data streams. This paper presents a data property categorization scheme associated with sensors used for monitoring the environment, and it describes how the categorization facilitates data property specification and supports improved data quality.

Dempster-Shafer Based Information Quality Processing in Smart Environments
Soukaina Messaoudi, Kamilia Messaoudi, and Serhan Dagtas

Abstract: Smart environments refer to buildings or locations equipped with a multitude of sensors and processing mechanisms for improved security, efficiency or functionality. Often, these sensors serve distinct purposes and their data may be processed separately by entirely separate systems. We argue that integrated processing of data available from multiple types of sensors can benefit a variety of decision making processes. For example, smart building sensors such as occupancy or temperature sensors used for lighting or heating efficiency can benefit the security system, or vice versa. Recent industry standards in sensor networks such as ZigBee make it possible to collect and aggregate data from multiple, heterogeneous sensors efficiently. However, integrated information processing with a diverse set of sensor data is still a challenge. We provide means to use a data fusion scheme that offers efficient processing of information collected from multiple sensors such as temperature sensors or motion detectors and visual sensors such as security cameras. The broader goal of multi-sensor data fusion in this context is to enhance security systems, improve energy efficiency by supporting the decision making process based on relevant and accurate information gathered from different sensors. In particular, we investigate the use of Dempster-Shafer based data fusion model and present techniques for processing of visual sensor data to facilitate the use of Dempster-Shafer model.

Research Presentations

Parallel Session 3-A Saturday, November 13, 2010 3:30 pm to 4:30 pm	Web Session Chair: Robert Pokorny EIT 224

Information Quality Challenges in Social Media

Nitin Agarwal and Yusuf Yiliyasi

Abstract: Social media has become an integral part of people's lives. People share their daily activities, experiences, interests, and opinions on social networking websites, opening the floodgates of information that can be analyzed by marketers as well as consumers. However, low barriers to publication and easy-to-use interactive interfaces have contributed to various information quality (IQ) problems in the social media that has made obtaining timely, accurate and relevant information a challenge. Approaches such as data mining and machine learning have only begun to address these challenges. Social media has its own distinct characteristics that warrant specialized approaches. In this paper, we study the unique characteristics of social media and address how existing methods fall short in mitigating the IQ issues it faces. Despite being extensively studied, IQ theories have yet to be embraced in tackling IQ challenges in social media. We redefine social media challenges as IQ challenges. We propose an IQ and Total Data Quality Management (TDQM) approach to the Social media challenges. We map the IQ dimensions, social media categories, social media challenges, and IQ tools in order to bridge the gap between the IQ framework and its application in addressing IQ challenges in social media.

A Framework for Developing Better Instruments to Measure Web-Based Health Information Quality

Samar I. Swaid

Abstract: With the Internet becoming a growing source of health information, evaluating the quality of the web-based health information is essential. A number of consumer-oriented instruments were designed to assess the quality of health information. However, existent tools are incomprehensive, unvalidated, not easy to use, and difficult to understand. The main objective of this study is to provide a framework for systematically developing validated and usable instruments to evaluate the quality of web-based health information. The framework incorporates item-to-total correlation and factor analysis. Next, weighting scores for the quality criteria are assigned based on the application of the Kano model to classify quality criteria as basic, performance or attractive. The development process ends with testing the instrument in terms of its usability, readability, reliability and validity. Among the implications and contributions of this research is improving the quality of the evaluation instruments and making the process of developing assessment tools more standardized and transparent.

Research Presentations

Parallel Session 3-B Saturday, November 13, 2010 3:30 pm to 4:30 pm	IQ Models I Session Chair: Beverly Kahn EIT 218

Information Quality for Operations: Framework and Model
Zbigniew J Gackowski

Abstract: A framework and model of information quality anchored in operations-management theory and based on realistic assumptions. It shows the role of information as a factor in operations, how theories of operations management affect information needs with their respective quality requirements, and how information qualitatively and quantitatively affects models; it identifies the universally necessary use requirements that make information usable and useful at four levels of expectations to make operations effective, ethical, efficient, or both ethical and efficient. Requirements are objectively categorized, ordered, and prioritized for examination in research and practice. It proves the futility of aggregated direct measures of quality and proposes an aggregate that indirectly measures how information, its quality aspects, and properties contribute to the extent that the purpose of operations can be attained.

Using Information Capacity to Assess Integrated Schema
Andrea Maurino, Carlo Batini, and Simone Grega

Abstract: In large organizations the database architecture is typically built through a series of projects and realizations that result in a number of heterogeneous and overlapping data sources. This trend is worsened by merger and acquisition activities that add new data sources from external organizations to the existing data architecture. Data fragmentation significantly reduces the possibility for an organization to exploit its information assets. A technology that partially alleviates these problems is data integration middleware that allows users to read-only access data stored in heterogeneous data sources through the presentation of a unified view of these data. In this paper we introduce a new concept to measure the improvement in information capacity enabled by data integration solutions, and propose an original framework which can support the evolution of the organization data architecture by identifying the optimal solution that maximizes such improvement within a given cost threshold.

Research Presentations

Parallel Session 3-C **Saturday, November 13, 2010** **3:30 pm to 4:30 pm**	**Invited Speaker II** **Session** **Christian Walenta** **EIT 220**

The Job of the Information / Data Professional

Christian Walenta, President of IAIDQ and Head of the IQ Certification Program

Abstract: For the first time, a broad group of information/data quality practitioners across the globe has reached consensus on the definition of the job of information/data quality professional. This achievement represents a major milestone in the establishment of information/data quality as a distinct profession. This breakthrough is the outcome of a job analysis study recently conducted by IAIDQ. This presentation describes the consensus definition of the job of an Information Quality Professional, organized into a framework containing six performance domains, twenty-nine tasks, and more than three hundred distinct knowledge areas and skills. After it was validated by a large international group of information/data quality practitioners, the framework was used to develop the specifications for IAIDQ's upcoming CIQP exam. Beyond this primary purpose, the framework is also expected to drive an increase in the quality and consistency of the information/data quality training available in the market place, and to provide a benchmark against which organizations can assess their information/data quality practices

Research Presentations

Parallel Session 4-A Sunday, November 14, 2010 9:30 am to 10:30 am	Data Migration Session Chair: Lwanga Yonke Chico Room

Data Quality as a Key Success Factor for Migration Projects

Dephine Clement, Soumaya Ben Hassine Guetari, and Brigitte Laboisse

Abstract: Organizations are facing the challenges of data migration, which results either from mergers and acquisitions due to the consolidation of the economy, or from upgrade of systems or finally from willingness to simplify the data storage architecture. A data migration is not a trivial task; either, it is a significant project undertaking. Because the data stored in the legacy systems is a strategic company asset, it needs to be analyzed, measured, preserved and improved before being brought over to the target system. In this paper, we present business cases and best practices on how companies structure their data quality management in the particular context of data migration. We propose a comparison grid with classified items such as the project management teams, the data quality audit and the key quality indicators (KQIs), the data quality improvement activities, the optimum quality level calculation, the problem of the data references and the need for a data governance post migration.

Migrating and Cleaning Data Using Excel: A Case Study

John Wilton and Anne Matheus

Abstract: This is a case study from a senior project in an undergraduate Information Systems program. It involves a real agency that could not afford to have the project completed and requested help from a small liberal arts college in their community. This case study is only one part of a larger project that included the development of a relational database and the creation of a web based application. The problems associated with the migration of the old data into the new system seemed insurmountable, but the challenge was accepted and overcome with excellent results using Excel.

Research Presentations

Parallel Session 4-B Sunday, November 14, 2010 9:30 am to 10:30 am	IQ Models II Session Chair: Farhad Moeeni Grampas Room

A Discussion of Methods for the Detection of Errors in a Power Law Distribution
Valerie Sessions and Chris Nuhn

Abstract: To detect erroneous data points within a set, one must first understand how the data is usually distributed. Often in our analysis of data, we assume a data set to be normal (that is to follow a normal distribution or be normally distributed). This often is not the case, however, particularly when dealing with human-centric events such as web site hits, word usage, or author citations. In this paper, we examine the power law distribution and the types of data that should be tested for such a distribution. We present test methods and results from the testing of the robustness of a power law distribution under conditions of randomly generated errors in the data set, and discuss methods for error detection in data sets with this type of distribution. We conclude with a discussion of future research.

A General Ranking Strategy for Data Accuracy Management
Irit Askira Gelman

Abstract: A series of recent studies proposed a construct named damage and a set of models for estimating the damage in a chosen class of information systems. The perception that underlies the proposed construct is that, all other things being equal, it would be beneficial to assign priority to the elimination of data errors that have a stronger negative effect on output accuracy (i.e., output accuracy is lower) over data errors that have a weaker effect. In this paper we extend the work on damage by considering its use with information systems in general, rather than a specific class of information systems. Mainly, we propose a general strategy for ranking the inputs according to the damage that errors in each input inflict on the output of the system. A major advantage of this strategy is that it focuses the ranking effort on a subcomponent of the information system that can be substantially smaller and simpler than the information system as a whole.

Research Presentations

Parallel Session 4-C Sunday, November 14, 2010 9:30 am to 10:30 am	IQ Tools Development Session Chair: Andrea Maurino Ouachita Room

Snap-On Data Quality Enhancement and Verification Tool (DEVA) For Asset Management
Jing Gao and Andy Koronios

Abstract: During the data collection process, human error is a large reason asset management organizations suffer from poor data quality (specifically, as not all data entries can be replaced by the automatic acquisition method). Thus, to reduce human error as one cause of data quality problems, software assistance is considered valuable. This paper provides an innovative client-server based software solution. By using the operating system interception method, this solution can transfer the data quality (business) rules generated from the data sets and apply them to the existing information system at the client side real-time. This snap-on approach has great potential to be applied in domains other than asset management, resulting in great enhancement of data quality.

A New Cycle of Improvement for Information Quality Services
Sandra Collovini de Abreu, Tércio Oliveira de Almeida, Luiz Carlos Ribeiro Junior, and Marilia Terra de Mello

Abstract: Poor quality information produced/used by organizations impacts directly on these companies' revenues. Several losses caused by quality problems could be minimized if data quality solutions were adopted. However, due to specific needs of each organization, the process of deploying these solutions is extremely costly. Some efforts to minimize deployment costs are being made by using Supervised Machine Learning techniques. However, these approaches demand template datasets which can be a drawback to their adoption. In this paper, we present and evaluate a synthetic address data generator aiming at minimizing this problem, focusing on a new cycle of improvement for information quality services. Our proposal covers an Address Validator service, Optimizer (based on the Particle Swarm Optimization) and the data generator. The complete approach can also be generalized to other domains. In a previous work we presented the evaluation of our training process (Address Validator and Optimizer). Here we assess our synthetic address data generator, closing the proposed cycle.

Research Presentations

Parallel Session 5-A Sunday, November 14, 2010 10:45 am to 11:45 am	IQ Fundamentals Session Chair: Ningning Wu Chico Room

Framing Data Quality Research: A Semantic Analysis Approach

Roger Blake and G. Shankaranarayanan

Abstract: Research in data and information quality has made significant strides in the last decade and has created an expansive body of knowledge. Given the multiple different research perspectives and research methodologies adopted, it is important for us to understand the research topics and themes that have evolved and currently define this body of research. Here, we present the results of a preliminary study that aims to provide a better understanding of this research area by identifying the core topics and themes. We analyze abstracts of 467 journal and conference articles published over the past ten years in data and information quality. Latent semantic analysis (LSA) is used to develop term-to-term semantic similarities and term-to-factor loadings. From the analysis, we identify five core topics and fourteen core themes of data quality research. The results from this research can significantly improve our understanding of the body of literature in data and information quality. Taken a step further, this research can offer insights into how themes and topics have shifted over time, what topics/themes have garnered the attention of researchers and when, and reveal the research trends in this area. Above all, it can motivate research by helping researchers associate research methods with research topics, identify themes/topics that have not been studied and data quality dimensions that have not been examined sufficiently.

Analysis of Information Quality Criteria in a Crisis Situation as a Characteristic of Complex Situations

Therese Friberg, Stephan Prödel, and Rainer Koch

Abstract: In the past a lot of researchers have defined criteria to determine information quality. Various criteria and dimensions have been identified and examined in different contexts. But very few of them focused on information quality in the context of complex situations, especially in the domain of crisis management. Characteristics of this domain include the facts that the stakes are high and the time for intervention is short. These complex situations demand for an extensive level of information as a basis to the difficult decisions an officer-in-charge has to make. But the need for quick action limits the time available for comprehensive information. Therefore, if we want to support the decision-making of an officer-in-charge through an at least semi-automated process, we need first of all to find a set of criteria to assess the information quality considering the special requirements of such complex situations. In this paper we describe our approach of defining a criteria set by identifying the characteristics of complex situations first, then analyzing existing models of information quality and mapping their aggregated criteria to the obtained characteristics and finally the preparation of a survey to evaluate the set through the involvement of domain experts.

Research Presentations

Parallel Session 5-B Sunday, November 14, 2010 10:45 am to 11:45 am	IQ Models III Session Chair: Mariofanna Milanova Grampas Room

The Effect of Data Quality on Data Mining – Improving Prediction Accuracy by Generic Data Cleansing

Jørgen Stang,, Tore Hartvigsen, and Joachim Reitan

Abstract: The recent advent and abundance of mainstream tools for performing business intelligence such as data mining and predictive analysis have made applications previously limited to highly trained specialists accessible to a wide range of users. At the same time, the amount of available data is rapidly increasing due to several factors such as government regulations, affordable hardware, mature master data and data warehouse sources and available 3rd party data for areas like demographics and financial analysis. As a consequence, advanced data mining techniques for predictive analysis find increasingly novel uses ranging from detecting mechanical failures in propeller shafts to forecasting customer retention. The usefulness of the data mining models will be highly dependent on the quality of the underlying data source and misrepresented data could distort the analysis and produce erroneous predictions which at best will be useless and at worst jeopardize operations. The data flaws could be hard to detect by manual inspections of the data mining results as these are frequently non intuitive. Therefore, the data quality should be carefully evaluated prior to implementing any data mining models and any findings should either be cleaned or used as an indicator for the confidence level of the resulting data mining model. This will in turn establish the effect of the data quality on the data mining as described in [7]. Also, the reverse approach have been suggested, where data mining is used as a successful means to detect data quality indicators [11] [16]. This article describes a practical case which attempt to assess the effect of improving the data quality on the prediction accuracy for a work order database.

Managing Accuracy of Project Data in a Distributed Project Setting

G. Shankaranarayanan, Nitin R. Joglekar, and Edward G. Anderson Jr.

Abstract: Organizations (principals) manage projects by outsourcing tasks to partners. Coordinating and managing such projects requires sharing project-data, status data on the work-in-progress residing with the partners and estimates of completion time. Project data is rarely accurate due to errors in estimation, errors in aggregating data across partners and projects, and gaming by the partners. While managers are aware of the inaccuracies, they are forced to make decisions regarding outsourcing the tasks (how much, to whom, and when). In this paper, we develop a control theoretic model that analyzes utilization of capacity of both the principal and partners. This model also permits corruption of project-data regarding progress status. We use this model to compute the costs of using perfect project-data versus inaccurate project-data and show that these costs can be significant. We propose a control policy, using filters, to correct inaccurate project-data and generate an estimate of true progress status. We show that this policy, depending on the relative magnitude of variances in demand and measurement error, can minimize cost-penalties due to inaccurate project-data.

Research Presentations

Parallel Session 5-C **Sunday, November 14, 2010** **10:45 am to 11:45 am**	**High Performance Computing Session** **Chair: Valerie Sessions** **Ouachita Room**

Towards a High Performance Merge Solutions for Large-Scale Datasets
Marília T. de Mello, Gustavo C. Frainer, José G. C. de Souza, and Luiz C. R. Junior

Abstract: Solving the Merge & Purge problem is a crucial step for Data Quality applications since they depend on integrated views of entities. In the present work, we discuss a domain-independent solution to the problem of detecting duplicate records and its implementation in a real-world application. Our approach is based on the Standard Blocking strategy with some enhancements proposed to overcome the major challenge of de-duplication in the industry scenario: maintaining good accuracy while processing large-scale datasets in a timely manner. Experimental results are given on both synthetic and real-world datasets. The performed experiments show that, with the proper configuration, it is possible to achieve high accuracy results, reaching an F-measure of 99.78%. Moreover, the performance evaluation shows that the proposed solution can be applied to large-scale real-world datasets, since it processed a dataset with 45 million records in about 3 hours in a simple PC, and a dataset containing 156 million records in about 5 hours and 40 minutes in a more robust machine.

Towards a Data-Intensive Approach to Named Entity Recognition
O. Isaac Osesina and John R. Talburt

Abstract: Many difficult natural language and machine learning problems are now yielding to data-intensive solutions made possible by the advent of economical, high-performance computing. In data-intensive computing, complex rules systems and statistical models can often be replaced by the ability to scan large volumes of data to find exact or similar instances in which the solution is known. This paper discusses the problem of named entity recognition in unstructured textual information, a discussion of how data-intensive computing methods have evolved to address this problem, and a comparison of results obtained between different data-intensive methods.

Poster Presentations – On Display in the EIT Lobby

Factors of the Online Healthcare Information Quality: The Perspective of the African American Consumers

Samar Swaid, D'Andrea Magee, Veronica Maples, and Pakeshia Benton

Abstract: Large number of studies investigated the concept and metrics of online healthcare information quality to enhance consumers' health education and health promotion. However, a limited number of studies focused on the perception of minorities in evaluating the quality of healthcare information. A two-phase study was conducted to understand how African American consumers evaluate the information quality of healthcare websites. In phase one; we applied the think-a-loud protocol as qualitative research method to elicit the perspective of healthcare information seekers. The second phase, the Kano model (Kano et al., 1984) was used to classify the quality metrics onto the "must be" criteria, the "performance" criteria and the "excitement" criteria. The think-a-loud protocol resulted in identifying the following attributes of healthcare information quality: reliability, credibility, currency, relevance, sources, readability, external links, statement of purpose, authorship, navigation tools, search engines and website accessibility. In the second phase, a sample of 102 subjects answered our questionnaire to explain consumer's satisfaction with information quality and her/his dissatisfaction using the Kano model. Data analysis indicates that quality criteria of reliability, currency, relevance and readability are the "must" criteria that its absence will result in dissatisfaction. On the other hand, "performance" criteria that increase satisfaction are website usability, navigation tools, search engines and website accessibility. Criteria of external links, statement of purpose and authorship are found to be the "excitement" ones that exceed consumers' needs and expectations.

An Analysis of World Oil Reserves Data from an Information Quality Perspective

Yusuf Yiliyasi and Daniel Berleant

Abstract: In this research, we focus on the problem of determining total world oil reserves through an information quality (IQ) and probabilistic approach. For each country-by-country oil reserve datum, we apply an IQ framework to measure the quality of critical IQ dimensions. We then combine the qualities to obtain the overall IQ of the source datum. This is then translated into a probability distribution that represents error in the nominal reserve figure. Next, we predict the total world oil reserves by summing the distributions of the constituent countries. We hypothesize that the IQ of the world oil reserves will be higher than the average IQ of the constituent oil reserves data from different countries, based on the intuition that errors in constituent figures will tend to cancel out when added together. The key concept in quantifying this is the coefficient of variation, which is a function of mean and standard deviation (or alternatively, mean and variance). By using the same approach, the next phase of our research will address the more complex problem of determining the date of world peak oil (or the Hubbert Peak), the point at which world oil production reaches its maximum for all of time.

Mixed Entity Resolution on Web Documents

Ingyu Lee and Byung-Won On

Abstract: In the World Wide Web, Mixed Entity Resolution is a common problem when non-unique values are used as an identifier due to their homonym and polysemy. For example, if only surname is used as an identifier, one cannot distinguish "Vanessa Bush" from "George Bush." To resolve such mixed entities on the Web, we proposed a hybrid approach in this paper. We developed a method, named numClus, which estimating the number of clusters in the given data set. Then, we combined numClus with several supervised methods including K-means, Spectral Bisection, Latent Dirichlet Allocation, and Similarity Propagation. Experimental results show that our approach shows 16% improvement in accuracy.

Poster Presentations – On Display in the EIT Lobby

Classifying Detection Using Classifier and Feature Selection by Ant Colony Optimization: A Methodology for Identifying Approximately Duplicated Records

Jianjun Cao, Xingchun Diao, Mingchao Tan, Yi Du, and Jianmei Zhang

Abstract: In this paper, a new methodology for identifying approximately duplicate records is proposed. Identifying approximately duplicate records is regarded as a two-class problem. Similarity of three kinds of typical attributes is defined and normalized. They are character string type, enumeration type and date type. The similarity vector, which is composed of attribute similarity of two records, is the input of the classifier. Then, A new graph-based ant system with step-down action for subset problems is also proposed. A multi-object mathematical model for optimizing recall rate, precision and the size of attribute set is reprensented. The model is transformed to a single object model, and an arithmetic based on the ant system with step-down action is designed for the single object model. At last, the methodology is tested using ED (Euclid Distance) and SVM (Support Vector Machine) classifier, and its effectiveness is validated.

Quality of Health Data: Improve the Quality of Health Care

Sanjay Gupta and Sushil K. Meher

Abstract: The quality of health data has become increasingly important for providing quality health care. In health care field in India the patient load is very high. If we look at the statistics, some startling facts stare at us. India, a country of 1.3 billion people (as reported on March 10, 2008), According to the Medical Council of India, the allopathic doctor-population ratio at present works out to 1:1722. At AIIMS the annual load of the outdoor patients is 1,000,000 and capturing patient data in high volume of patient load is a big challenge for an organization. Patient's health data is very much essential for the effective health care and the accuracy of diagnoses and the practices of health care providers is crucial to organizations that strive to maintain and improve health care delivery. Hospital, clinics and other health care facilities are also under increasing pressure to treat more patients and give quality of health care. The term "data quality "can best be defined as "fitness for use" which implies the concept of data quality is relative. Thus data with quality considered appropriate for one use may not possess sufficient quality for another use. What kind of standardization is been implemented during the designing and implementing HIMS in an organization. The purpose of this paper is to examine the issue of health care organization to deliver high quality of patient's health information and make a model to capture the quality of clinical patient data.

Poster Presentations – On Display in the EIT Lobby

Review of Data Scrubbing and Cleaning Methods for Improving Information Quality
Richard S. Segall and Farhad Moeeni

Abstract: This paper provides a summary of research currently in progress for a more-in-depth study of available methods for improving information quality. The first part of the research-in-progress is a preliminary literature review of work by other researches of dealing with dirty data. This paper also provides an overview of preliminary research in locating available data sets of dirty data, and discuses future directions of the research. A multitude of diverse applications of data scrubbing to dirty data are discussed. We plan to expand this research to devise an expanded literature review and tutorial on data cleansing methods and techniques.

Using Entity Resolution to Discover Scientific Numbers
Shen Lu , Richard Segall, and Samin Atiff

Abstract: In this paper, we try to extract not only words but also numbers associated with words. However, there are so many numbers that the value of the variation cannot be extracted by using probability models. In this paper, we provide a method to discover number-word sequences from a text. Conditional Random Field model is not only a probability model, but also a framework for entity extraction. It can label words with both the probability of the word and the probability of the transition among the adjacent words. In this way, the frequent word sequences can be identified. Since numbers are infinity, no matter how big the training set is, we always have un-labeled data. Therefore, we used not only words but also the format of the word as features to segment information. Experimental result showed us that our solution effectively eliminate the problem of the un-labeled data. The ability to determine the named entities in a text has been established as an important application for several natural language processing areas, including information retrieval, machine translation, information extraction and language understanding. In this paper, we tried to extract entities about scientific numbers.

Consideration on Data Quality Dimensions for Long-Term Biological Observation of EcoSystems
Dongxiu Wu, Wenshan Wei, Chuangye Song, Ying Su

Abstract: Data quality is the life of monitoring working. The Chinese Ecosystem Research Network (CERN) was established in 1988 under the auspices of the Chinese government and the World Bank Loan. Through years of effort, it is now well placed to address important issues, serving as a functional network to meet the needs of both the national and international ecological research. The costs of making incorrect scientific inferences based on faulty data can be substantial and far-reaching, and follow-on research may be critically jeopardized. To improve scientific data quality, and provide continuous quality assessment and management, the nature of scientific data and the processes that produce it must be articulated. Although firms are improving data quality with practical approaches and tools, their improvement efforts tend to focus primarily on accuracy, consistency and completeness, with no clear DQ framework and these dimensions having no clear description and measuring method. The purpose of this research is to provide a framework for the management of data quality as it applies to scientific data, specifically those generated by the fieldwork facilities and instrumentation that will populate the data centers of CERN, based on data quality theory. Definitions for data and data quality tailored to the context of ecological research are proposed. This paper develops a list of data quality dimensions that captures the aspects of data quality which are important for ecosystem long-term monitoring.

Conference Venue & Location Maps

About UALR

The University of Arkansas at Little Rock (UALR) was founded in 1927 as Little Rock Junior College. That first semester there were eight instructors and about 100 students. By 1929 the college was accredited by the North Central Association of Colleges and Schools, a status it has kept through changes in size and stature.

Housed at first in public school buildings, the college moved in 1949 to its present location in southwest Little Rock on a beautifully wooded site donated by Raymond Rebsamen, a Little Rock businessman. The college was also by that time the sole beneficiary of a continuing trust established by former Governor George W. Donaghey. In 1957, the institution began a four-year degree program, became independent and privately supported under a separate board of trustees, and took the name Little Rock University. In September 1969, Little Rock University merged with the University of Arkansas to create the University of Arkansas at Little Rock.

Today UALR is a metropolitan university on the move, a dynamic learning institution where students find unique learning and internship opportunities provided through one-of-a-kind connections to the state's thriving capital city. With more than 100 programs of study, UALR has an academic program to suit many interests – and an equal number of social and service organizations as well. UALR is:

- A public, metropolitan university with 13,000-plus full and part-time students.
- Six colleges and the Bowen School of Law that include more than 100 programs of study.
- Small classes with a 14 to 1 student-to-faculty ratio.
- Part of the University of Arkansas System since 1969.
- Students come from all over Arkansas, 43 other states, and 62 foreign countries.
- The tree-lined campus has 40 buildings on 150 acres – and is still growing!
- Specialized facilities include a learning resource center, art galleries, radio station, television station, Cyber Café, wireless network, and speech and hearing clinic.
- The library contains 400,000 volumes, 700,000 microform items, 8,300 audiovisual forms and CDs, subscribes to 2,625 periodicals, and its computerized services include card catalog, interlibrary loan, and database searching.

To learn more about UALR, check out the website: http://ualr.edu/.

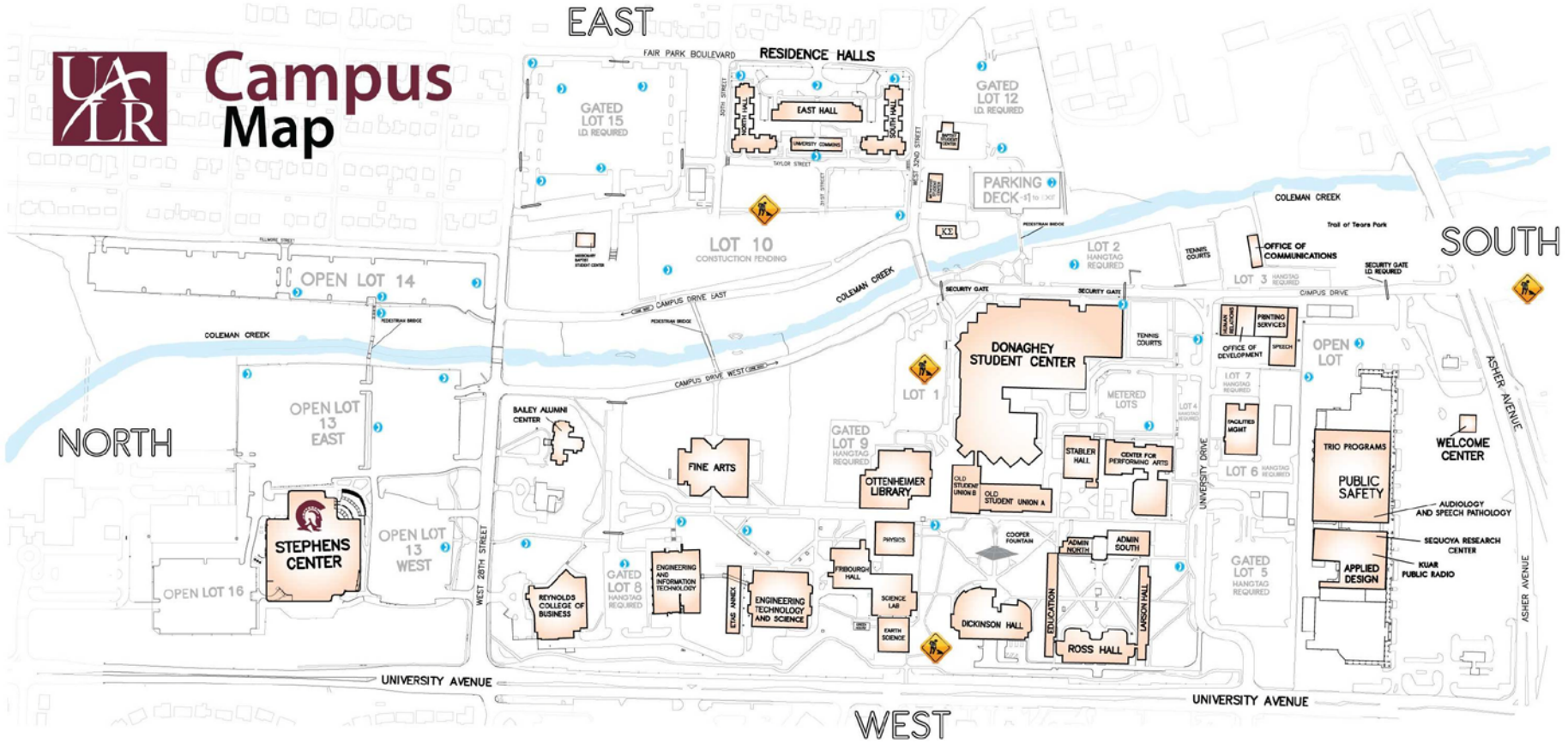
About the Donaghey College of Engineering and Information Technology (EIT)

The mission of the Donaghey College of Engineering and Information Technology is to educate the next generation of high-tech professionals in the skills and knowledge base necessary to create and manage the technology based enterprises that will provide future economic growth and an improved standard of living for the state of Arkansas and its citizens. The College's expectation is for every Arkansas child to have the opportunity to participate in the new knowledge-based digital economy of the 21st Century. This mission includes technological education at all levels, from high school through advanced graduate degrees, as well as contributions through scholarly research and community involvement.

In meeting this mission, the College offers professional undergraduate BS degrees in Computer Science, Construction Management, Construction Engineering, Engineering Technology, and Information Science. In addition, it offers a BS degree in Systems Engineering with options in Mechanical, Electrical, Computer and Telecommunications Engineering. EIT offers the only interdisciplinary doctoral degree in the state, the PhD in Applied Science, with options in six disciplines including Applied Computing and Engineering Science and Systems. Graduate students also have the option to pursue a PhD in Integrated Computing. The College participates in the MS-PhD. Program in Bioinformatics offered jointly with the University of Arkansas for Medical Sciences. Additional graduate programs in the College include MS degrees in Applied Science, Computer Science and Information Quality. It also offers a graduate certificate program and a MS degree in Systems Engineering. Extensive outreach to the general undergraduate population is through the College’s Computer Literacy classes and the acclaimed Information Technology (IT) Minor, designed to provide the nontechnology majors with the IT tools necessary to command leadership positions in today's IT-enabled enterprises.

Outreach to the community includes the IT certificate program for in-service learning and extensive partnering with high schools across the state for in-school activities and summer programs. Specific emphasis is on partnerships with local and regional industries ranging from direct company input into EIT’s programs to in-service courses and directed research projects.

UALR Campus Map



Important Places

Admissions and Financial Aid are located in the Admin South building.

The Donaghey Student Center (DSC) is home to:

- Dining Services / Cafeteria
- Fitness Center and Aquatics
- Bookstore
- Health Services

Additional dining locations can be found in the Reynolds Business building, the Engineeing and Information Technology building, and in the University Commons near student housing.

Computer Labs open to all students are located in Ottenheimer Library, and additional labs operated by various departments are located throughout campus.

Pardon Our Progress

Be prepared to encounter building construction zones in these areas:

North of the Donaghey Student Center, where a **One-Stop Student Services Center** is being built to house the most commonly visited administrative offices in one convenient location.

Behind Fribourgh Hall and adjacent to ETAS, under construction is a five-story **Center for Integrative Bio Nanotechnology Science** (CIBNS)

A **fourth Residence Hall** is being constructed near the others, designed to be a part of the residence village bordering on Fair Park Avenue. The six-story building will provide housing for approximately 370 students.

On the south of campus across Asher Avenue, under construction is a **Recreation and Sports Complex** to include track and field/soccer facilities.

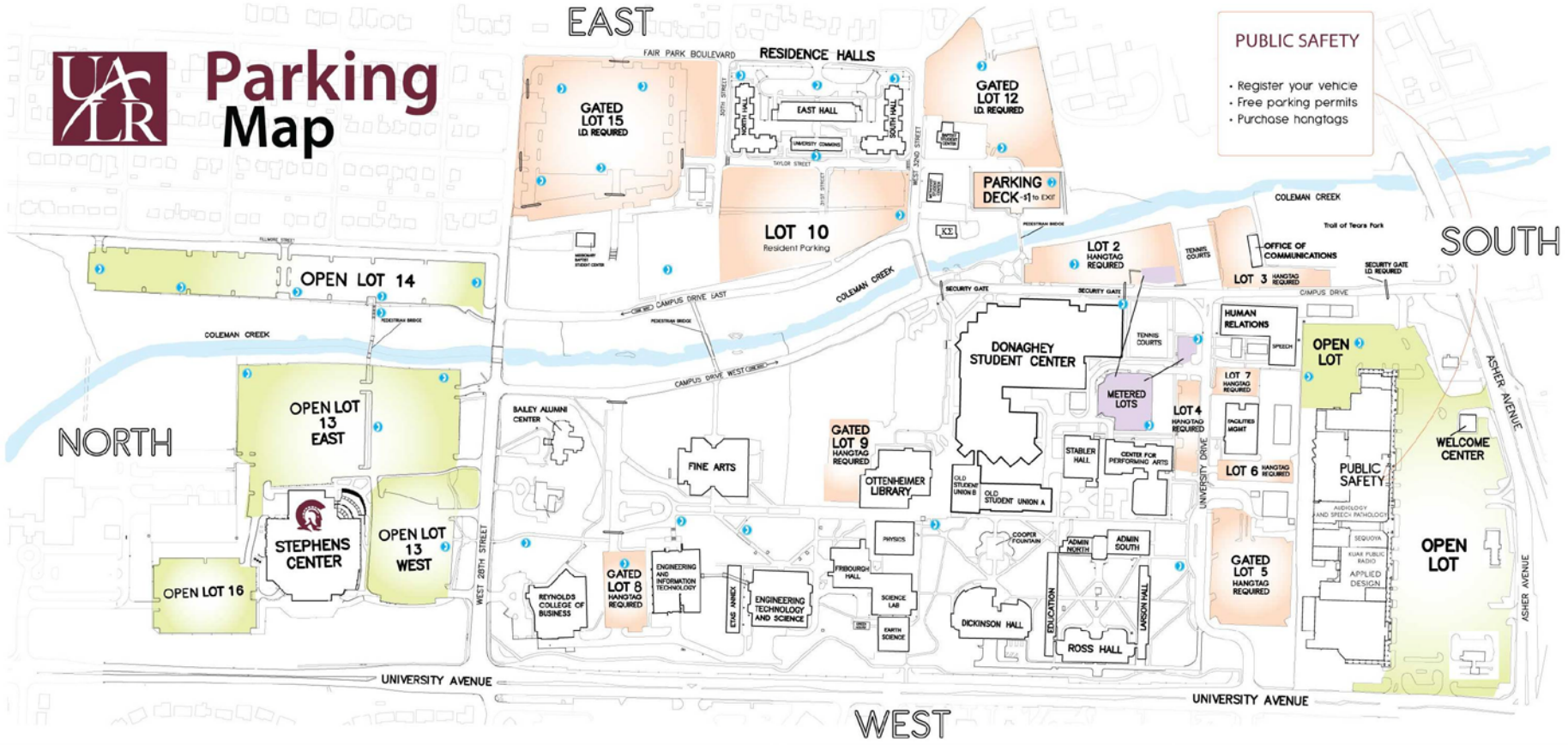
Help!

Are you lost? Do you need directions to a specific location on campus?
Call **501-569-3000**

Additionally, you can visit the University web site at ualr.edu and use the A-Z Index to directly contact an Office / Department / Individual.

Have you found a problem on this map? Do you have a suggestion to improve it?
Let us know. Send an email to maps@ualr.edu

UALR Parking Map



Parking Regulations

All vehicles must display a valid permit, or hangtag, and be registered with the Public Safety Department.

In addition to Open Lots, students are also allowed to park in the metered lots or UALR's parking deck. The fee for parking in the meter lots is $1 per hour with a 2 hour time limit, and the fee for the parking deck is $1 per exit.

A complete list of parking regulations can be found online: http://ualr.edu/publicsafety/TrafficRules.pdf

Parking tickets may be paid at the Cashier's Office, located in the Administration North building.

Visitor Parking

Visitors to UALR may park in any of the metered parking lots for a fee of $1.00 per hour with a 2 hour limit, or they may park in the parking deck for a fee of $1.00 per exit.

Night Parking

Any faculty member, staff member, or student can park in a gated lot after 4 p.m. by swiping their UALR picture ID card in the proper slot at the gate.

Handicapped Parking

Handicapped Parking spaces are available in every campus lot. Special arrangements for reserved handicapped parking can be made by contacting UALR Health Services.

Gates and I.D. Cards

Gated entrances at Asher Avenue and for Parking Lot 12, (the open parking lot on Fair Park and 32nd Street) will respond to a valid UALR identification card at any time of day or night.

Additionally, select gated lots will only open for those who have reserved parking in those lots.

To obtain an I.D. card, visit Reception Services in the Donaghey Student Center, located on the bottom floor by the main entrance of the gym, near the bookstore. Alternate forms of identification are required when applying for the UALR I.D. card.

In Case of Emergency

Contact the Office of Public Safety at (501) 569-3400

Trained security guards patrol campus, and they are available to escort students and others to campus destinations.

To request a security guard, call 569-3400, or use one of the blue emergency telephones designated on the map.


One pleasant trip with flying kangaroos
One taste of world top wines
One hug with koalas
One walk into the sunset
One valuable conf. in IQ
ICIQ 2011
One unique experience in Adelaide, Australia
UniSA

16th International Conference on Information Quality

Hosted by the University of South Australia

To be held in Adelaide, South Australia 18th to 20th of November 2011

Call for Papers

Theme: '*Disproportionate Organisational Value through Quality Information'*

Information quality is critical to the success of businesses. High quality information also gives companies a clear competitive advantage, whereas poor information quality management can lead to inaccurate forecasting, increasing costs, reputational problems, compliance issues and in extreme cases injury or loss of life

Conference Profile

The International Conference on Information Quality (ICIQ) was first held at MIT in 1996 by one of the pioneers of information quality Professor Richard Wang. The conference has, until 2009 been held annually at MIT and is the premiere conference in information quality. It not only attracts the international researchers from Universities and research institutions worldwide but also a large number of practitioners from the public and private sectors.

In 2009, it was for the first time, held outside MIT, at Potsdam in Germany and it is the first time to be hosted in Australasia during 2011. The University of South Australia is proud to be jointly sponsoring this prestigious conference with the MIT Information Quality Program. The International Conference on Information Quality (ICIQ brings together researchers and practitioners from the academic, private and public sectors to exchange their knowledge and ideas about information quality.

The University of South Australia is delighted to offer you this excellent opportunity for your organisation to be noticed by hundreds of researchers and practitioners in Australia and around the world. Furthermore if you decide to accept a sponsorship opportunity before the 10th of November 2010, you will be able to receive further exposure during this year's conference in the US.

Important Dates

May 10th , 2011: Submission deadline
June 3rd , 2011: Notification of acceptance
August 31st , 2011: Early registration deadline
September 6th , 2011: Camera-ready copy due
November 18, 2011: Conference starts

TOPICS OF INTEREST

Corporate and organizational IQ

- IQ Management and Data Governance
- Alignment of IQ with Business Strategies
- Business Process Performance
- IQ Education and Curriculum Development
- Methods, concepts, and tools for IQ
- IQ Concepts, Metrics, Measures, and Models
- Method Engineering for IQ
- Trust, Knowledge, and Society
- Data Provenance and Annotation
- Information Product Theory and Practice
- Metadata and IQ
- IQ of Unstructured and Extracted Data
- IQ in Probabilistic, Fuzzy, and Uncertain Data
- IQ in Sensor Networks and Information Fusion
- Data Quality Systems and Tools

Measurement, improvement, and assurance of IQ

- Data Scrubbing and Cleaning
- Record Linkage and Entity Resolution
- IQ Assessment, Policies, and Standards
- Cost/Benefit Analysis of IQ Improvement
- Privacy & Security Issues in Data Cleansing

IQ cases and applications

- IQ Practices: Case Studies and Experience Reports
- IQ in the Web, Data Integration, and e-Business
- Data Warehousing and Business Intelligence
- Master Data Management
- Scientific Data
- Community Input, Pay as You Go, and Crowd Sourcing

Conference Chairs

Andy Koronios, University of South Australia, Adelaide (andy.koronios@unisa.edu.au)

Jing Gao, University of South Australia, Adelaide (jing.gao@unisa.edu.au)

John R. Talburt, University of Arkansas at Little Rock (jrtalburt@ualr.edu)

Notes

Notes

Notes

Notes

Notes

Many thanks to the following sponsors for their support of ICIQ 2010!