ICIQ 2013

The 18th International Conference on Information Quality

Co-located with the
IAIDQ Information Data Quality (IDQ) Conference *and*
UALR Emerging Analytics Center

November 7-9
Donaghey College of Engineering & IT
University of Little Rock
Little Rock, Arkansas

Many thanks to the following sponsors for their support!

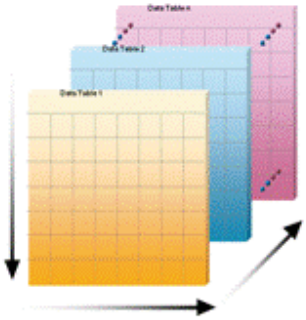
acxiom

IBM

TARGET

BlackOak PARTNERS

THE CENTER FOR IDENTITY
THE UNIVERSITY OF TEXAS AT AUSTIN



Data Profiling LLC

Mechdyne
ENABLING DISCOVERY

iaidq

UALR George W. Donaghey Emerging Analytics Center
University of Arkansas at Little Rock

ERIQ
Entity Resolution & Information Quality

Table of Contents

Welcome

Dear Conference Participants,

Welcome to the 18th International Conference on Information Quality (ICIQ 2013). 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 diverse IQ topics as frameworks, dimensions, risk management, entity resolution, organizational issues, tools, and case studies from healthcare, business, and big data applications. ICIQ 2013 also features two keynote talks and ample opportunities for informal discussion and networking. In addition the new Emerging Analytics Center at UALR will be sponsoring additional sessions on the role of information quality in advanced data analytics, innovations in information visualization, and new venture opportunities.

This is the second time for the University of Arkansas at Little Rock (UALR) to host ICIQ. Our campus is located on the western side of Little Rock, the capital city of the state of Arkansas. 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.

Our state capital of Little Rock features a downtown bustling River Market District with restaurants, shops, museums and hotels. Several focal points in the area are the Clinton Presidential Center & Park, the neighboring world headquarters of Heifer International and the Central Arkansas Nature Center. The arts flourish in Little Rock with the Arkansas Arts Center, home to one of the most-acclaimed collections of works on paper in the country. The city's historic sites include Little Rock Central High School National Historic Site, Historic Arkansas Museum, Old State House Museum, Mount Holly Cemetery, and MacArthur Museum of Military History. In addition, Little Rock's position in the center of the state 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 2013 conference and your visit to Little Rock!

John Talburt and Elizabeth Pierce
University of Arkansas at Little Rock

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. Graduate Certificate in Information Quality,
2. Master of Science in Information Quality, or
3. PhD in Integrated Computing with an Emphasis in Information Quality.

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

CONFERENCE SCHEDULE

Thursday, November 7 (Downtown Little Rock)			
5:30-7:00 pm	**Opening Reception and Registration sponsored by Acixom at the Acxiom River Market Tower located at 601 East 3rd Street, Little Rock, AR 72201**		
Friday, November 8 (EIT Building on the UALR Campus, 2801 South University Avenue, Little Rock, AR 72204)			
Shuttle Service	**Free daily bus transcription from the downtown hotels to the UALR Campus. Bus departs from the Little Rock Marriott – 3 Statehouse Plaza (7:45 am and 8:35 am).**		
8:00-9:00 am	**Registration, Morning Refreshments, EIT Lobby**		
9:00-10:30 am	**ICIQ 2013 Opening Ceremonies in the EIT Auditorium** • Welcome by Dr. Zulma Toro, UALR Provost • Recognition of Sponsors and EAC Advisory Committee • Reports by IQ Societies, Invitation to ICIQ 2014 • Keynote Speaker: Dr. Scott Schumacher, IBM Distinguished Engineer and Chief Scientist for IBM InfoSphere MDM, "Entity Resolution and Data Quality Frameworks for Big Data"		
10:30-10:45 am	**Coffee Break**		
	EIT 220	**EIT 219**	**EIT 419**
10:45-12:15 pm Parallel Sessions 1A and 1B	IQ and Entity Resolution, Session IA Chair: Valerie Sessions • *A False Positive Review Indicator for Entity Resolution Systems Using Boolean Rules* • *On Choosing Thresholds for Duplicate Detections* • *An Approach Using Relational Markov Model for Estimating and Replacing Missing Categorical Data*	IQ in Organizations, Session IB Chair: C. Lwanga Yonke • *An Industry Comparison of Data and Information Quality Priorities and Practices* • *Does the Organization Size Matter? An Investigation into IQ Effort in AIS Adoption* • *Modeling and Simulating the Impact of Social Issues on Information Quality: Literature Review*	Emerging Data Analytics Panel, Session 1C / Led by Dr. Mary L. Good, Special Advisor to the Chancellor for Economic Development, University of Arkansas at Little Rock • *The Emerging Synergy between Data Quality and Data Visualization with three topic categories featuring guests from the UALR EAC Data Science Advisory Board*
12:15-1:15 pm	Box Lunch on EIT Second Floor.		
	EIT 220	**EIT 219**	**EIT 419**
1:15-2:45 pm Parallel Sessions 2A and 2B	IQ of Wireless Sensor Networks, Session 2A Chair: Michael Mielke • *Big Data Quality Case Study Preliminary Findings: HSI Using The AVIRIS* • *Semantic-based Detection of Segment Outliers and Unusual Events for Wireless Sensor Networks*	IQ in Healthcare, Session 2B Chair: Bruce Davidson • *Monitoring the Quality of Clinical Administrative Health Data to Support Health System Funding* • *Information Quality and Data Governance for eHealth in the Era of Big Data*	Innovation in Emerging Data Visualization, Session 2C / Demos hosted by EAC Technical Director Dr. Edi Tudoreanu • *Medical Data Visualization featuring UAMS Dr. T. Glen Pait (invited)* • *Other Special Data Visualization Applications*

CONFERENCE SCHEDULE

Friday, November 8 (UALR Campus, EIT Building)			
2:45-3:00 pm	**Coffee Break**	**2:45-3:00 pm**	**Coffee Break**
	EIT 220	**EIT 219**	**EIT 419**
3:00-4:30 pm Parallel Sessions 3A and 3B	IQ and Risk Management, Session 3A Chair: Ismael Caballero • *Total Information Risk Management* • *Methods for Adjusting Expected Value of Information Under Situations of Data Missing Not at Random* • *Show Us Your Pay Stub: Income Verification in P2P Lending*	Healthcare and IQ Issues Panel Session 3B Chair: Beverly Kahn • *Bruce Davidson* • *Alein Chun* • *Sami Laine* • *Meredith Nahm*	Emerging Ventures in Data, Session 3C led by moderator Lee Watson / Founder, Clarovista, Co-Founder Southern Coding, Co-Founder, Arkansas Venture Center *Session to include invited guests representing the Arkansas Venture Center leadership team and the UALR College of Business Data Analytics Development Committee*
4:30-6:00 pm	**Opening Day Reception sponsored by Mechdyne at the UALR Bailey Alumni Center**		
Shuttle Service	**Departs UALR Campus for Downtown Hotels (Little Rock Marriott – 3 Statehouse Plaza at 4:45 pm and 6:15 pm)**		

CONFERENCE SCHEDULE

Saturday, November 9 (UALR Campus, EIT Building)			
Shuttle Service	**Free daily bus transcription from the downtown hotels to the UALR Campus. Bus departs from the Little Rock Marriott – 3 Statehouse Plaza (7:45 am and 8:35 am).**		
8:00-9:00 am	**Registration, Morning Refreshments, EIT Lobby**		
9:00-10:30 am	**ICIQ 2013 Morning Ceremonies in the EIT Auditorium** • Welcome Back, Awards and Announcements • Keynote Speaker: Dr. Carolina Cruz-Neira, W. Hansen Hall and Mary Officer Hall / Board of Regents Super chair in Telecommunications, Electrical and Computer Engineering, University of Louisiana at Lafayette, "Data Around Me: Immersive Explorations to Turn Data into Information"		
10:30-10:45 am	**Coffee Break**		
	EIT 220	**EIT 219**	
10:45-12:15 pm Parallel Sessions 4A and 4B	Tools for IQ, Session 4A Chair: Jing Gao • *Data Profiling Challenges in Engineering Asset Management Data – Conceptual Design for Next Generation Data Profiling Software* • *Solution Architectures for Retaining Data Quality Problems in Automatically Generated Test Data* • *Systematic ETL Management – Experiences with High-Level Operators*	Case Studies IQ, Session 4B Chair: Mariofanna Milanova • *A Study of the Promotion of Information Sharing: Case Study of Tabio Corp. in Japan* • *The Application of the IQMM Model to Evaluating the Science and Technology Information Resources Sharing Projects in China* • *Improving Customer Complaint Data Mining*	
12:15-1:15 pm	Box Lunch on EIT Second Floor		
	EIT 220	**EIT 219**	
1:15-2:45 pm Parallel Sessions 5A and 5B	IQ Dimensions, Session 5A Chair: John Talburt • *Perception of Value-Added Through a Visual Join Operation* • *User Interaction Metadata for Improved Information Traceability*	IQ Frameworks, Session 5B Chair: Elizabeth Pierce • *18K: An Implementation of ISO 8000-1X0* • *The Information Value Methodology: Web User Survey Preliminary Results* • *Research on Information Quality Viewed By Praxiology*	
2:45-3:00 pm	**Coffee Break**		
	EIT 220	**EIT 219**	**EIT 218**
3:00-4:30 pm	IQ and Super Computing Workshop	Emerging Research Challenges in IQ and Big Data Workshop	Emerging Markets and Collaborations for IQ Workshop
Shuttle Service	**Departs UALR Campus for Downtown Hotels (Little Rock Marriott – 3 Statehouse Plaza at 1:00 pm, 3:00, 4:45 and 5:40 pm)**		
6:00 – 8:00 pm	**Survivor's Event: Dinner for Those Interested at a local Little Rock Restaurant**		

ICIQ 2013 Organization

Conference Chairs
John R. Talburt, University of Arkansas at Little Rock, USA
Wei Huang, Xi'an Jiaotong University, Xi'an, China
Jacky Akoka, Conservatoire National de Arts et Métiers, Paris, Franc

Program Chairs
Elizabeth Pierce, University of Arkansas at Little Rock, USA
Bruce Davidson, Hoag Health System, Newport Beach, CA, USA
Michael Mielke, Deutsche Bahn AG, Frankfurt, German

Program Reviewers
Elizabeth Pierce, University of Arkansas at Little Rock
Bruce Davidson, Hoag Health System
David Becker, MITRE Corporation, USA
Laure Berti-Equille, Institut de Recherche pour le Developpement, France
Ismael Caballero, University of Castilla-La Mancha, Spain
Cinzia Cappiello, Politecnico di Milano, Italy
InduShobha Chengalur-Smith, SUNY at Albany, USA
Tamraparni Dasu, AT&T Labs – Research, USA
Adir Even, Ben-Gurion University of the Negev, Israel
Craig Fisher, Marist College, USA
Jing Gao, University of South Australia, Australia
Michael Gertz, Ruprecht-Karls-University Heidelberg, Germany
Cesar Guerra, UPSLP, Spain
Olaf Hartig, University of Waterloo, Canada
Markus Helfert, Dublin City University, Ireland
Beverly Kahn, Suffolk University, USA
Stephen Kennet, Australia Department of Defense – DSTO, Australia
Barbara Klein, University of Michigan at Dearborn, USA
Hiroshi Koga, Kansai University, Japan
Andy Koronios, University of South Australia, Adelaide, Australia
Brigitte Laboisse, Global Data Excellence, France
Eitel Lauria, Marist College, USA
Peggy Leonowich-Graham, United States Military Academy, USA
Philip Lesslar, IAIDQ, Malaysia
Jiuyong Li, University of South Australia, Australia
Bart Longenecker, University of South Alabama, USA
Piyush Malik, IBM, USA
Helina Melkas, Lappeenranta University of Technology, Finland
Mariofanna Milanova, University of Arkansas at Little Rock, USA
Paolo Missier, The University of Manchester, UK
Felix Naumann, Hasso-Plattner-Institut an der Universitat Potsdam, Germany
Paulo Jorge Oliveira, Politécnico do Porto, Portugal
Barbara Pernici, Politecnico di Milano, Italy

Leo Pipino, University of Massachusetts Lowell, USA
Robert Pokorny, XSB, Inc., USA
Srini Ramaswamy, ABB India Corporate Research, India
Thomas Redman, Navesink, USA
Shazia Sadiq, The University of Queensland, Australia
Monica Scannapieco, Italian National Institute of Statistics (Istat), Italy
Thomas A. J. Schweiger, Acxiom, USA
Laura Sebastion-Coleman, Ingenix, USA
Yasuki Sekiguchi, Hokkaido University, Japan
Valerie Sessions, Charleston Southern University, USA
Ganesan Shankaranarayanan, Babson College, USA
John (Skip) Slone, Lockheed Martin Corp., USA
Ying Su, Institute of Scientific and Technical Information of China, China
Giri Kumar Tayi, SUNY at Albany, USA
Mihail E. Tudoreanu, University of Arkansas at Little Rock, USA
Niels Weigel, SAP AG, Germany
Rolf Wigand, University of Arkansas at Little Rock, USA
Philip Woodall, Cambridge University, UK
Harris Wu, Old Dominion University, USA
Ningning Wu, University of Arkansas at Little Rock, USA
C. Lwanga Yonke, Aera Energy LLC, USA

Planning and Advisory Committee
John R. Talburt, University of Arkansas at Little Rock, USA
Elizabeth Pierce, University of Arkansas at Little Rock, USA
Candice High, University of Arkansas at Little Rock, USA
Kim Tran, University of Arkansas at Little Rock, USA
Melody Penning, University of Arkansas at Little Rock, USA
Therese Williams, University of Arkansas at Little Rock, USA
Thomas Wallace, University of Arkansas at Little Rock, USA
Stephanie Boccarossa, University of Arkansas at Little Rock, USA
Ashekul Huq, University of Arkansas at Little Rock, USA

Keynote Talks

Dr. Scott Schumacher

Initiate Chief Scientist, Information Management, IBM

Topic: Entity Resolution and Data Quality Frameworks for Big Data

ICIQ 2013 Friday Keynote Speaker
November 8, 2013

Abstract: In today's competitive world, enterprises seek to exploit new data sources that can provide insight into their customers' needs and interests. These insights provide micro-segmentation capabilities to marketing efforts, support customer centricity programs, allow more accurate next best action (NBA) predictive models, and enable many other customer-oriented initiatives. Cost-effective processing and storage systems, such as Hadoop, provide the enterprise with the means to maintain and analyze these data, such as click-stream data, deep purchasing history, customer interaction history, and social media data.

However, in this big data world, many of the standard data quality and entity resolution techniques, long used for processing data internal to the enterprise, either aren't applicable, don't scale, or are too manually intensive. In this talk we explore some new techniques being developed for entity resolution and data quality evaluation specific to the big data platform and discuss how they are being applied to customer problems.

Speaker Biography: Scott Schumacher, Ph.D., is a well-known technology expert specializing in statistical matching algorithms for healthcare, enterprise, and public sector solutions. For more than 20 years Dr. Schumacher has been heavily involved in research, development, testing, and implementation of complex data analysis solutions, including work commissioned by the Department of Defense.

As chief scientist, Scott is responsible for the research and development of the Initiate matching algorithms, and holds multiple patents in the entity resolution area. He is also responsible for the Initiate Master Data Service product architecture.

Scott has a Bachelor of Science degree in Mathematics from the University of California, Davis, and received his Master of Arts and Doctorate degrees in Mathematics from the University of California, Los Angeles (UCLA). He is currently a member of the Institute for Mathematical Statistics and the American Statistical Association.

Keynote Talks

Carolina Cruz-Neira

W. Hansen Hall and Mary Officer Hall / Board of Regents Super chair in Telecommunications, Electrical and Computer Engineering, University of Louisiana at Lafayette

Topic: Data Around Me: Immersive Explorations to Turn Data into Information

ICIQ 2013 Saturday Keynote Speaker
November 9, 2013

Abstract: We are living in an exciting period in human history where we can describe, quantify, qualify, and simulate almost everything about the world we live in, and we are doing that through large amounts of data. But as we accumulate all this valuable data, we are also facing the challenge of how to extract meaningful information within reasonable time frames for the issues under investigation. Conventional approaches to data analysis and visualization do not work in the data sets that are being generated today. We need an innovative, fresh, and "out there" approach that combines the best of human capabilities and the best of computers' abilities to extract information in a manageable time frame and targeted to the intended consumers. Part of the new approach is to realize that both humans and computers have complimentary capabilities suitable for data analysis and the challenge is to understand what those capabilities are and how to orchestrate them to best take advantage of both. Immersive environments provide an excellent stage to bring humans and computers into a common ground and work together towards discovering the information that hides in the big data.

Certainly we are living the "next era of computing" and it is indeed exciting. This talk explores the use immersive technologies integrated with the human domain expert knowledge and intuition to gain insight from large data and provides, where are the challenges, the research opportunities, and the type of teams that would be successful in this endeavor. The talk also includes an overview of several projects Dr. Cruz is currently involved in this topic as well as her lessons learned through the years.

Speaker Biography: Dr. Carolina Cruz-Neira is the W. Hansen Hall and Mary Officer Hall/BORSF Endowed Super Chair in Telecommunications in Computer Engineering. Dr. Cruz was the founding Executive Director (CEO) of the Louisiana Immersive Technologies Enterprise (LITE), a State of Louisiana initiative to support economic development through the use of immersive technologies. She was also the LITE's Chief Scientist and a member of the Board of the Greater Lafayette Chamber of Commerce until 2012.

Prior to being in Louisiana, Dr. Cruz was the Stanley Chair in Interdisciplinary Engineering and the Associate Director and co-founder of the Virtual Reality Applications Center at Iowa State University (ISU). In 2002, she founded and co-directed the Human-Computer Interaction

graduate program at ISU. Dr. Cruz's work in VR started with her PhD dissertation, the design of the CAVE Virtual Reality Environment, the CAVE Library software specifications and implementation and preliminary research on CAVE Supercomputing integration. She is known as the co-inventor of the CAVE and the original developer of the CAVELibs. Since then, her research is driven by providing applicability and simplicity to VR technology focusing on software engineering for VR, applications of VR technology and usability studies of virtual environments. She spearheaded the open-source VR API movement with the development of VR Juggler and has been an advocate of best practices on how to build and run VR facilities and applications. She has chaired several international conferences, given over fifty keynote addresses, serves in a number of review boards for National and International funding agencies, and participates in technology advisory task forces for the U.S. Federal and State Governments defining the research directions of her field. Many of her former students are now doing leading work in VR at places such as Purdue University, Navteq, Nintendo, EA, Deere & Company, Boeing, Sony Pictures Imageworks, the U.S Navy and Army, and Argonne National Laboratory.

Beyond her academic career, Dr. Cruz is a business entrepreneur. She co-founded Glass House Studio, a company dedicated to create virtual experiences and she also co-founded Infiscape Corporation, a services company in immersive applications and high-end interactive graphics. She serves on many advisory boards, including Sensics Inc., Mersive and Micoy and has performed corporate consulting for many companies around the world. She has also designed and produced stage performances and public exhibits in New York, Chicago, Los Angeles, Orlando, Tokyo, Madrid, Barcelona, Florence and other places combining technology, dance, theater, and art.

Among her many achievements, in 1997, Business Week magazine named Dr. Cruz a "rising research star" in the new generation of computer science pioneers. In 2001 she received the Boeing A.D. Welliver Award, in 2002 she was inducted as Eminent Engineer by the Tau Beta Pi Honors Society, in 2003 she was inducted as a Computer Graphics Pioneer by the ACM SIGGRAPH organization, in 2007 she was the recipient of the Virtual Reality Technical Achievement Award from the IEEE Visualization and Graphics Technical Committee (VGTC), and in 2009 she received the International Digital Media and the Arts Association Distinguished Career Award.

Dr. Cruz has a PhD in Electrical Engineering and Computer Science (EECS) from the University of Illinois at Chicago (UIC) (1995) and a master's degree in EECS at UIC (1991). She graduated cum laude in Systems Engineering at the Universidad Metropolitana at Caracas, Venezuela in 1987. Outside her professional interests, Dr. Cruz was an accomplished Classical Ballet Dancer until 1990. She also does artistic creations involving 3D technology, and her work has been shown in a number of Art museums around the world. A set of her art pieces is in permanent exhibit at the Museum of Jewish Heritage in New York City. She enjoys traveling around the world and she is very active in a number of animal rescue organizations.

Research Presentations

Parallel Session 1-A: Entity Resolution and IQ **Friday: November 8, 2013** **10:45 am to 12:15 pm**	**EIT 220** **Session Chair: Valerie Sessions**

A False Positive Review Indicator For Entity Resolution Systems Using Boolean Rules

Daniel Pullen, Pei Wang, John Talburt, and Ningning Wu

Abstract: The clerical review of potential false positive and false negative resolution decisions is critical to improving the accuracy of an entity resolution (ER) process. In ER systems using scoring rules or agreement/disagreement patterns, review indicators are triggered as pairs of references are compared. In systems using scoring rules match pairs that may need clerical review can be indicated by match scores falling within a particular value range. For systems using agreement/disagreement patterns match pairs that satisfy a particular agreement/disagreement pattern are selected for review. ER systems using Boolean match rules require a different approach. This paper describes a new method for identifying potential false positive resolutions made by an ER process based on the entropy of the identity attribute values across the complete set of references that have been linked together by the process. The method has the advantage that it can be applied to any ER process outcome regardless of the type of match rules used by the process. The method is efficient in identifying false positives in large data sets and has been implemented and tested for student enrollment data. The paper also discusses an analysis for estimating the precision and recall of false positive resolutions for various entropy value thresholds.

On Choosing Thresholds for Duplicate Detection

Uwe Draisbach and Felix Naumann

Abstract: Duplicate detection, i.e., the discovery of records that refer to the same real-world entity, is a task that usually depends on multiple input parameters by an expert. Most notably, an expert must specify some similarity measure and some threshold that declares duplicity for record pairs if their similarity surpasses it. Both are typically developed in a trial-and-error based manner with a given (sample) dataset.

We posit that the similarity measure largely depends on the nature of the data and its contained errors that cause the duplicates, but that the threshold largely depends on the size of the dataset it was tested on. In consequence, configurations of duplicate detection runs work well on the test dataset, but perform worse if the size of the dataset changes. This weakness is due to the transitive nature of duplicity: In larger datasets transitivity can cause more records to enter a duplicate cluster than intended. We analyze this interesting effect extensively on four popular test datasets using different duplicate detection algorithms and report on our observations.

An Approach Using Relational Markov Model For Estimating and Replacing Missing Categorical Data

Jianjun Cao, Xingchun Diao, Yongping Xu, and Zhen Yuan

Abstract: In order to process missing data, we propose an approach based on the relational Markov model (RMM) for estimating and replacing missing categorical data. First, for a given data set, all categorical attributes are classified as a proper number of groups, and these groups are independent of each other. Second, principles for ordering attributes in one group are proposed and the attribute sequence of the group could be indexed by the principles. Third, a RMM for estimating missing categorical value is represented. According to complete record samples, probabilities of missing value belonging to each possible value are estimated by the model. The dynamic attribute selection (DAS) method is used to select the best attribute group to estimate the missing value. The missing values could be replaced using maximum posterior probability (MaxPost) or probability proportional (ProProp) method. Finally, the effectiveness and advantage of the approach is tested by the comparative experiments on open datasets.

Research Presentations

Parallel Session 1-B: IQ in Organizations **Friday: November 8, 2013** **10:45 am to 12:15 pm**	**EIT 219** **Session Chair: C. Lwanga Yonke**

An Industry Comparison Of Data And Information Quality Priorities And Practices
Elizabeth Pierce and Bruce N. Davidson

Abstract: This work expands upon the analysis conducted for the joint IAIDQ-UALR IQ 2012 Industry Survey and Report on the State of Information and Data Quality [1]. In this paper we examine the differences and similarities between industries when it comes to priorities, managerial approaches, practices, tools, and maturity levels for information and data quality efforts. These comparisons raise interesting questions about how information and data quality efforts are influenced by the type of industry associated with an enterprise.

Does The Organization Size Matter? An Investigation Into IQ Effort in Accounting Information Systems Adoption
Andy Koronios, Wongsim Manirath, and Jing Gao

Abstract: Many organizations have undertaken information quality improvement programs and projects. In order for an organization to better target their IQ efforts, this research has conducted 10 case studies to study how IQ is managed through the accounting information system adoption process. A special focus is placed on determining how organization size influences the information quality practices. The finding is especially useful to Small and Medium Enterprises (SMEs) as many SMEs have the desire to grow bigger. By better dealing with IQ issues, there could be a successful future.

Modeling and Simulating the Impact of Social Issues on Information Quality: Literature Review
Therese L. Williams, David Becker, Thomas C. Redman, Amit Saha, Kashif Mehdi, Joanne Reilley, Huzaifa Syed, Wright A. Nodine, Jr., and John Talburt

Abstract: Information quality is generally defined in terms of fitness for use. Almost all agree that they prefer high-quality to low-quality information. And, while many organizations have made good progress, many find that setting up information quality programs and making improvements proves difficult. Further, most agree that the most critical difficulties stem from organizational, structural and political issues. As yet, there is no body of theory and practice to help leaders and organizations systematically understand and address these issues.

This research program aims to (begin to) build the body of needed theory. The basic idea is to employ systems dynamics and computer simulation to explore the ways hundreds of possible factors and managerial actions advance or hinder information quality efforts. More specifically then, the long-term goal of this research is to create and utilize a test bed (or simulator) to examine, in a systematic fashion, the impact of various social/cultural issues which influence the penetration and overall success of information quality in an organization. In particular, building on the work of Falleta (1), this research is a literature review of multiple organizational change models that can potentially be utilized for this modeling.

This paper reports on one aspect of this research, namely, the literature review. As one might suspect, there is much relevant work, from the fields of systems dynamics, organizational analyses, force-field analyses, and change management.

Research Presentations

Parallel Session 2-A: IQ of Wireless Sensor Networks **Friday: November 8, 2013** **1:15 pm to 2:45 pm**	**EIT 220** **Session Chair: Michael Mielke**

Big Data Quality Case Study Preliminary Findings: Hyperspectral Imaging (HSI) Using The Airborne Visible / Infrared Imaging Spectrometer (AVIRIS)
Dave Becker, Trish Dunn King, Bill McMullen, and Dr. Ahmed Fahsi

Abstract: In this study, we examined Big Data Quality issues using the case of the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS), hyperspectral imaging (HSI) sensor. The study addresses several factors affecting Big Data Quality at multiple levels, including collection, processing, and storage. Though not unexpected, the key findings of this study reinforce that the primary factors affecting Big Data reside in the limitations and complexities involved with handling Big Data while maintaining its integrity. For example, with the HSI imaging data, the composite, end-to-end scientific data pipeline interactions can affect its data quality. These concerns are of a higher magnitude than the provenance of the data, the processing, and the tools used to prepare, manipulate, and store the data.

Semantic-based Detection of Segment Outliers and Unusual Events for Wireless Sensor Networks
Lianli Gao, Michael Bruenig, and Jane Hunter

Abstract: Environmental scientists have increasingly been deploying wireless sensor networks to capture valuable data that measures and records precise information about our environment. One of the major challenges associated with wireless sensor networks is the quality of the data – and more specifically the detection of segment outliers and unusual events. Most previous research has focused on detecting outliers that are errors that are caused by unreliable sensors and sensor nodes. However, there is an urgent need for the development of new tools capable of identifying, tagging and visualizing erroneous segment outliers and unusual events from sensor data streams. In this paper, we present a SOUE-Detector (Segment Outlier and Unusual Event-Detector) system for wireless sensor networks that combines statistical analyses using Dynamic Time Warping (DTW) with domain expert knowledge (captured via an ontology and semantic inferencing rules). The resulting Web portal enables scientist to efficiently search across a collection of wireless sensor data streams and identify, retrieve and display segment outliers (both erroneous and genuine) within the data streams. In this paper, we firstly describe the detection algorithms, the implementation details and the functionality of the SOUE-Detector system. Secondly we evaluate our approach using data that comprises sensor observations collected from a sensor network deployed in the Springbrook National Park in Queensland, Australia. The experimental results show that the SOUE-Detector can efficiently detect segment outliers and unusual events with high levels of precision and recall.

Research Presentations

Parallel Session 2-B: IQ in Healthcare **Friday: November 8, 2013** **1:15 pm to 2:45 pm**	**EIT 219** **Session Chair: Bruce Davidson**

Monitoring the Quality of Clinical Administrative Health Data to Support Health System Funding

Maureen Kelly and Chrissy Willemse

Abstract: This presentation shares new data quality monitoring tools that were developed by the Canadian Institute for Health Information (CIHI) and the Ontario Ministry of Health and Long-Term Care to support the use of healthcare data for funding purposes.

Methods: Clinical administrative data are now being used to determine funding allocations to regions and hospitals across multiple health sectors using clinical administrative health databases maintained at CIHI. CIHI worked with the Ontario ministry to develop tools to monitor clinical data quality indicators for health care organizations on a quarterly basis, focusing on data that plays an important role in funding or that were found to be problematic through data analysis.

The tools provide a new way of looking at data quality. Data are examined at an aggregate level using graphical representation to enable quick identification of outliers. Charts automatically populate for a selected organization enabling comparisons with others and trending over time. Data visualization is used to highlight changes over time and hospital-level differences making it easy for users to identify potential data quality problems. Interpretation notes are included in the report to guide the user and encourage them to consider whether the data is an accurate representation of their patient populations and hospital processes. Questions are also posed for the user to consider whether any changes or anomalies in their data are a result of a data quality initiative or known problems.

Results and Discussion: These data quality monitoring tools foster a culture of data quality by enabling healthcare providers and health system managers to identify data quality problems and take action on an ongoing basis. This provides confidence to health system planners that the data are fit for use for funding and management purposes.

Information Quality and Data Governance for eHealth In the Era of Big Data

Ying Su, Cheng Dong, Marc Lange, Junpping Zhao, and Ping Yu

Abstract: This paper focuses on the second class of Big Data, and in particular on secondary uses of health care datasets in the National Health and Family Planning Commission (NHFPC) in China. For the purposes of discussion, the primary use of health care information is for the diagnosis, treatment and care that we receive from doctors, nurses and other clinicians. We ask two questions - why information quality is essential to the eHealth in the era of Big Data, and how to govern health big data? We suggest that health big data analytic engine and big data governance tool for high knowledge creation and personalized health service. We also conclude that, high quality health information about the user will lead to better personalized services, and better adaptive services.

Research Presentations

Parallel Session 3-A: IQ and Risk Management **Friday: November 8, 2013** **3:00 pm to 4:30 pm**	**EIT 220** **Session Chair: Ismael Caballero**

Total Information Quality Risk Management: Quantifying the Business Impact of Information Quality
Alexander Borek

Abstract: A vast body of empirical evidence has been collected which shows that using information of poor quality can cause significant risks in the organization. This practice- oriented talk presents a methodology for managing and reducing the risks inherent in information quality in form of a process for "Total Information Risk Management" (TIRM). The TIRM process has been developed in a rigorous design science cycle, which included the application of the process in six industrial in-depth case studies using participative and non-participative research and evaluation interviews with ten information management experts. The TIRM process was refined after each application using the feedback and insights collected in these studies. The studies have shown that a risk based approach can be used to quantify the business impact of information quality, which can drive the development of a robust business case for information quality initiatives.

Methods for Adjusting Expected Value of Information (EVPI) Under Situations of Data Missing Not at Random (MNAR)
Valerie Sessions and Stan Perrine

Abstract: Decision making under uncertainty is an extensive research field concerned with aiding the decision maker through uncertain problem spaces such as financial markets, product analysis, or medical treatment options. It is often helpful in this type of problem space to obtain additional data before making a risky or costly decision. The benefit of this new data, however, must be weighed against the cost of obtaining it. One can use the Expected Value of Perfect Information (EVPI) calculation to determine the expected payoff for receiving new information regarding a future decision. The EVPI calculation is used to place an upper bound on the amount of funding or other resources that should be spent to 'firm up' data related to an uncertain situation. Used traditionally in business analysis, this method is becoming more pervasive in medical research, education, and criminal justice. While EVPI is a useful estimate of data utility, we postulate that under cases of poor data quality, specifically data missing not at random (MNAR), EVPI calculations can be misleading without further clarification. We seek here to present the possible effects of data MNAR in the EVPI calculation. We then give examples of EVPI usage in medical literature and the reporting of sample size, missing data, and mitigation strategies used (such as list wise deletion or multiple imputation). Finally, we propose two methods that can be used to inform a decision maker of the possible effects of presumed MNAR data into the EVPI calculation. Future research shall focus on evaluating this method more fully in data sets where MNAR data is suspected.

Show Us Your Pay Stub: Income Verification In P2P Lending
Authors Irit Askira Gelman and Aimee A. Askira

Abstract: Peer-to-peer lending is an alternative credit market that allows individual borrowers and lenders—people like you and me—to engage in credit transactions without traditional banking intermediaries. This research centers on income verification practices in peer-to-peer lending. We report on a descriptive analysis of all the loans that were funded through Lending Club, currently the world's largest peer-to-peer lending platform, with issue dates before September 1, 2012. The score that Lending Club assigns to a requested loan is purported to encapsulate all the information that is needed for the lender to assess the risk of a potential default. This study points, however, to a potential weakness of Lending Club's loan assessment tools, which indicates that information about a loan's income verification status is in fact relevant and has value. Given this understanding, lenders' choices are surprising. Lenders that are registered directly on Lending Club's platform, including a crowd of small investors, fund a higher percentage of the listed loan amount when the borrower's income is not verified, while all other investors display the opposite, traditional risk-averse behavior.

Research Presentations

Parallel Session 3-B: Healthcare and IQ Round Table **Friday: November 8, 2013** **3:00 pm to 4:30 pm**	**EIT 219** **Session Chair: Beverly Kahn**

Healthcare and IQ Issues: A Round Table Discussion Led By

Dr. Bruce Davidson
Vice President, Performance Improvement, Hoag Health System

Dr. Alein Chun
Interim Director, Resource & Outcome Management at Cedars-Sinai Health System

Dr. Meredith Nahm
Associate Director for Clinical Research Informatics, DTMI Biomedical Informatics Core, Duke University

Mr. Sami Laine
Department of Computer Science and Engineering, Aalto University

Abstract: Information quality is an increasingly critical topic for the healthcare industry. This panel will lead an interactive discussion with participants on a variety of data quality issues associated with Electronic Health Records, Research Reporting, and Healthcare System Data. Questions to be discussed include:

1. Does the quality of healthcare information matter to clinical decisions, to facility management, and to secondary data users
2. What are all if the information uses and the IQ dimensions important to reach?
3. Will clinicians use information provided by / charted by others and under what conditions?
4. How can we make information processes transparent and how do we leverage this metadata?
5. What metadata is important in healthcare?
6. Can we improve information quality by standardizing data elements for data generate and used in care/wellness management, i.e., can the data elements be standardized, will they be accepted and will they be adopted in EHRs
7. How to assess quality, particularly accuracy of health IQ?

Research Presentations

Parallel Session 4-A: Tools for IQ **Saturday: November 9, 2013** **10:45 am to 12:15 pm**	**EIT 220** **Session Chair: Jing Gao**

Data Profiling Challenges in Engineering Asset Management Data – Conceptual Design for Next Generation Data Profiling Software
Jing Gao, Philip Woodall, Andy Koronios, and Ajith Kumar Parlikad

Abstract: Engineering asset management (EAM) is the process of managing the assets (from manufacturing machines to trains, planes and road bridges etc) in an organization. In order to manage these assets organizations must have good quality data about the assets. Otherwise, decisions about when to maintain an asset can be made incorrectly, and as a consequence, can adversely impact the business financially. To improve data, the first commonly accepted stage is data quality assessment, and to support this stage, data profiling software is often used. Data profiling tools can be used to uncover and measure the scale of the data quality problems and they do this by defining data quality rules. This research investigated the data profiling needs of EAM. In particular, existing profiling tools often contain generic data quality rules that are not always applicable to EAM business users. Creating EAM data quality rules without the relevant domain knowledge is very difficult and hence the best people to develop these rules are the EAM business users. This research therefore proposes an enhanced data profiling solution, which is based on the community-based central pseudo-code DQ rule repository. The proposed data profiling solution enables business users to develop and share EAM-related data quality rules promoting rule adaptability and reusability.

Solution Architectures For Retaining Data Quality Problems in Automatically Generated Test Data
Martin Oberhofer, Philip Woodall, and Alexander Borek

Abstract: Many organizations store sensitive data that should not be released. However, organizations often want to release this data to benefit from outsourcing of work or using the cloud for Data Quality (DQ) related tasks like data cleansing, for example. Our previous work identified useful “data generation” methods that can modify secret data to make it releasable and also retain the original DQ problems. However, there are many different types and uses of organizational data. Our aim for this work, therefore, is to present the specific architectures for data generation methods that are applicable to organizations for their different types and uses of data.

Systematic ETL Management – Experiences With High-Level Operators
Alexander Albrecht and Felix Naumann

Abstract: Large organizations load much of their data into data warehouses for subsequent querying, analysis, and data mining. Extract-Transform-Load (ETL) workflows populate those data warehouses with data from various data sources by specifying and executing a set of transformations forming a directed acyclic transformation graph (DAG). Over time, hundreds of individual ETL workflows evolve as new sources and new requirements are integrated continuously into the system. Managing these, often complex, ETL workflows is a daunting task.

We built an ETL management framework to improve this difficult task by providing high-level operations, such as searching, matching, or merging ETL workflows. In this paper, we present our lessons learned throughout the implementation of a prototypical ETL management framework. We discuss our observations and experiences and highlight selected suggestions and algorithms, which we propose to be suitable for building useful ETL management operators.

Research Presentations

Parallel Session 4-B: Case Studies in IQ **Saturday: November 9, 2013** **10:45 pm to 12:15 pm**	**EIT 219** **Session Chair: Mariofanna Milanova**

A Study of the Promotion of Information Sharing Through Presentation of Suppositional Context and Using A Concept of Corporate Household: Case Study of Tabio Corp in Japan
Hiroshi Koga

Abstract: The author proposes to use two concepts, suppositional context and corporate household, to study how the Tabio Corporation in Japan (TCJ), one of good players in the Japanese hosiery industry, improved quality of information related to POS (point of sales) data which is provided for its socks suppliers and raw material provider (The trading company which handles raw thread and dye factories) to those socks suppliers. Through a sociological analysis, the author first shows that TCJ improved the users' perspective on how to utilize the data provided (suppositional context), not by modifying the information system but by introducing the “voluntary delivery system” of socks to the distribution center. Second, the author shows that significance of entity integration (corporate household) from a view point of the socks suppliers. Finally, the author focuses on the system expansion made by TCJ to provide POS data of each shop for raw material provider. That is, the POS data was aggregated into quantities of users' products, then supplied to the users.

The Application of the IQMM Model to Evaluating the Science and Technology Information Resource Sharing Projects in China
Lirong Song and Xiaohong Zhang

Abstract: The problems associated with Information Quality (IQ) have become prominent restraints for China's efforts for promoting science and technology (S&T) information resource sharing. It is important to use comprehensive, authentic and accurate means to describe, measure, and evaluate the Information Quality Management (IQM) of S&T information resource sharing projects. In this paper, a framework of Information Quality Management Maturity (IQMM) assessment is proposed, which is focused on the maturity features, constituent elements, key links and processes and evaluation criteria and methods of maturity levels of IQMM to provide effective strategies and measures to improve IQ in the S&T information resource-sharing construction. Finally, the paper presents the preliminary results of a case study, where IQMM assessment is applied to some of the specific S&T information resource sharing management projects in the national program "National Science and Technology Foundation Platform" in China, to improve IQ continuously.

Improving Customer Complaint Mining
Matt Brown

Abstract: This presentation details the use of naïve Bayes classification to address information quality issues in a customer complaint database from a large food manufacturer. The complaint database is primarily used to identify potential problems in products, production facilities, or retail locations. The database contains information on over one million unique customer complaints and grows on average by several hundred complaints daily, as customers contact the company with complaints. The verbatim complaint is manually classified into one of numerous predetermined complaint categories. Based on this classification, the database is monitored for significant changes in the rate of complaints using Statistical Process Control (SPC) based data mining techniques. The information quality issue arises when call center operators, who often are pressed for time between calls, misclassify complaints into “unknown” or “miscellaneous categories”. Subsequently, incomplete or inaccurate data limits the effectiveness of the SPC data mining technique. This paper examines the effectiveness of naïve Bayes classification in correcting these misclassified complaints.

Research Presentations

Parallel Session 5-A: IQ Dimensions **Saturday: November 9, 2013** **1:15 pm to 2:45 pm**	**EIT 220** **Session Chair: John Talburt**

Perception Of Value-Added Through A Visual Join Operation
Ahmed Abuhalimeh, Daniel Pullen, and M. Eduard Tudoreanu

Abstract: The quality of data and information can be judged and improved via multiple dimensions, such as degree of accuracy, degree of uncertainty, or the amount of value-added. Value-added, the focus of this paper, is one of the contextual information quality dimensions that depends on the nature of task and plays a role in the operational fitness and the goals to be achieved as ascertained by an end-user, the final information consumer. This paper presents an empirical study of how people perceive the value-added of data undergoing a join operation, which is common in both data processing and visual information fusion. The study focuses on data that is conveyed to end-users through visual representations, because graphical approaches are increasingly employed to convey large amounts of information. Two types of visualizations are shown to users, one with the basic, unprocessed data, and the other with the result of the join operation. Results show that there is an actual value added by the join operation, and we estimate it at between 24% and 35% in the context of the experiment described. This study is part of related research on the human perception of information quality conveyed through graphical means.

User Interaction Metadata for Improved Information Traceability
Sami Laine, Marko Nieminen, and Mika Helenius

Abstract: The origin of information must be traceable to determine its true contextual meaning and actual quality. Information flows can be traced from two complementary perspectives: information management and software systems. In management practices, various modeling methods are used to document and analyze information flows. In software systems, data lineage capability provides automatic traceability of information products back to their original data sources. However, accuracy of information is affected by many contextual and human factors that are unrecognizable from technical data flows or abstract management models. In this study, we suggest that information traceability methods should cover more metadata on such factors.

We use design science research approach to gradually develop three iterations of artifacts: theoretical categorization, laboratory prototype, and software application. In the first iteration, we gathered empirical data about user interaction situations to illustrate the potential use cases for user interaction metadata. We also categorized user interaction metadata properties that could be collected automatically from computerized processes, workflows, and user interfaces. We mapped user interaction metadata to five categories: interaction context, documentation, input controls, value properties, and interaction structures. Results from empirical cases suggest that additional user interaction metadata could help to recognize, explain, and fix contextual data quality problems in real-life information production processes.

During later iterations, our research project will develop prototypes to collect user interaction metadata and then evaluate them in real organizational settings. In the future, software-based data lineage capabilities should expand their coverage to user interaction metadata. Also, managerial modeling and analysis tools could use user interaction metadata to discover, analyze, and model actual organizational processes in more detail.

Research Presentations

Parallel Session 5-B: IQ Frameworks **Saturday: November 9, 2013** **1:15 pm to 2:45 pm**	**EIT 219** **Session Chair: Elizabeth Pierce**

I8K: An Implementation of ISO 8000-1X0
Ismael Caballero, Isabel Bermejo, Luisa Parody, Mª Teresa Gómez López, Rafael M. Gasca, and Mario Piattini

Abstract: The exchange of master data between organizations can be regulated by the family of standards ISO 8000-1x0. This family of standards imposes several requirements for some of the activities derived from the used of master data management. These requirements are the existence of a specific syntax to be stored in a data dictionary (ISO 8000-110), the necessity to include information about the data provenance (ISO 8000-120), the necessity to include information about the assessment and certification of data quality levels for the dimensions of accuracy (ISO 8000-130) and completeness (ISO 8000-140).

The data exchanged between applications requesting data and their corresponding data providers tend to be developed by means of Web Services, and this way of communication is covered by the standard ISO 8000-1x0. In order to satisfy the requirements of the family of standards, two elements are included in this paper as the main contribution: The I8K architecture that that implements some of the requirements of the specified part of ISO 8000 with a web service approach., and the ICS-API, which provides developers with the necessary primitives to communicate the applications to use the standard with the I8K architecture. With the aim of illustrating the use of both components, we also introduce an example in the domain of travels, in which an application named TripPlanner exchanges data with a flight provider named FlightCIA.

The Information Value Methodology: How Average Users Assess IQ On The Web – Preliminary Results
Marilou Haines and Elizabeth Pierce

Abstract: The Internet is a self-regulating complex system where users decide what is important by their actions. Since the burden of locating and evaluating information depends on knowledge, experience and skill, this study investigates the web-user experience in a rigorous and holistic manner. The survey instrument, populated with vetted Quality Characteristics from key multi-disciplinary literature, best practices and international standards presents the point of view of academics and practitioners. Over 200 students and faculty of a large U.S. university assessed the criticality of those dimensions. The analysis of their responses will advance our knowledge on the missing factor: “the perspective of the information consumer” and serve as a baseline for future research. This paper offers preliminary results of the data collected through April 2013.

Research On Information Quality Viewed by Praxiology
Zbigniew J. Gackowski

Abstract: This paper discusses the current Framework for Information Quality Research from the praxiological perspective, the theory of human conduct; authors of the Framework encourage interdisciplinary approaches. The praxiological perspective reveals the need for extension of the topics of the framework with regard to how differently information quality is perceived along the line of command, when strategies and theories of operations management shift, and when one faces the dangers of disinformation, misinformation, or even outright information warfare as studied in war academies.

I8K: AN IMPLEMENTATION OF ISO 8000-1X0

(Research-in-Progress)

Ismael Caballero[1], Isabel Bermejo[1], Luisa Parody[2], Mª Teresa Gómez López[2], Rafael M. Gasca[2] and Mario Piattini[1]

[1]University of Castilla-La Mancha, Spain
{Ismael.Caballero, Isabel.Bermejo, Mario.Piattini}@uclm.es,

[2]University of Seville, Spain
{lparody, maytegomez, gasca}@us.es,

Abstract: The exchange of master data between organizations can be regulated by the family of standards ISO 8000-1x0. This family of standards imposes several requirements for some of the activities derived from the used of master data management. These requirements are the existence of a specific syntax to be stored in a data dictionary (ISO 8000-110), the necessity to include information about the data provenance (ISO 8000-120), the necessity to include information about the assessment and certification of data quality levels for the dimensions of accuracy (ISO 8000-130) and completeness (ISO 8000-140).
The data exchanged between applications requesting data and their corresponding data providers tend to be developed by means of Web Services, and this way of communication is covered by the standard ISO 8000-1x0. In order to satisfy the requirements of the family of standards, two elements are included in this paper as the main contribution: The I8K architecture that *that implements some of the requirements of the specified part of ISO 8000 with a web service approach.*, and the ICS-API, which provides developers with the necessary primitives to communicate the applications to use the standard with the I8K architecture. With the aim of illustrating the use of both components, we also introduce an example in the domain of travels, in which an application named TripPlanner exchange data with a flight provider named FlightCIA.

Key Words: ISO 8000-1x0, Master Data Management, Data Quality, Data Quality Certification

1. INTRODUCTION

Organizations need data to feed their organizational processes. As it is recognized, the success of these processes is grounded, among other factors, in the quality of the used data. This is why it can be considered that along with people, data is one of their most important assets. Once organizations recognize this fact, their managers tend to think that as much data as they can collect and store, the more efficient they will be, and consequently, the larger is the probability of succeed in an increasingly competitive market. Organizations manage instances of data which are considered as part of the domain in which they run their business. These data, commonly known as Master Data, are considered as critical for their business and essential for their business processes [1].

As known, organizations, to achieve their business goals, needs to run activities which use data; data which is commonly instantiated from master data. As the activities can interact with other elements outside of the borders of the organizations, the exchange of data between organizations will be necessary. The values of these data can correspond to different instances of master data, with high probability of having different implementations of the same master data for two different organizations. In order to make a successful exchange of data, and consequently, to achieve the business goals of the organizations, these organizations must raise a consensus about the meaning and format of the exchanged data to avoid different interpretations and implementations of the master data. This consensus will avoid situations as

using different names for representing the same thing, or describing the data with different types, for example when for an organization a data corresponding to "Name" is represented by means of *nvarchar2(50)* whereas for another one is represented as a *String*.

In order to make easier the exchange of versions of data corresponding to the same master data between organizations, the usage of tools and techniques provided by the Master Data Management foundations [2] can be largely helpful. Even more, it would be convenient to deal not only with Master Data Management concerns, but also with issues related to the level of quality of the data being exchanged. In this sense, the family of standard ISO/IEC 8000-1x0:2009 can help organizations since it provides an approach to the Master Data Management in the exchange of master data, with a special focus on data quality of the data being exchanged. The family of standards provides a set of requirements that organizations must follow in the master data exchanged messages. These requirements are: to implement a Data Dictionary [3] formatted according to a specific syntax containing the terms corresponding to master data and enabling the operations of codifying and de-codifying (ISO 8000-110:2009 [4]); to add information about the data provenance (ISO 8000-120:2009 [5]), to add information about the evaluation and certification of accuracy (ISO 8000-130:2009 [6]), and about evaluation and certification of completeness (ISO 8000-140:2009 [7]).

As part of our research, and being conscious of the advantages and benefits of implementing this family of standards and making them available to organizations, we analyzed these requirements. The first conclusions we reached was that, in order to make a usable implementation of the ISO 8000-1x0 family of standards, we needed to look at the requirements from the point of view of the developers who build the software applications that exchange data. This first conclusion led us also to focus the problem from a server/client perspective: on the one hand, it was necessary to develop a set of servers which implement the requirements imposed by the different parts of the standard; on the other hand, it was necessary to provide developers with the adequate mechanisms (like an API).

In order to take the advantages introduced by the standard, we propose: (1) to implement a solution supporting the previously stated requirements provided by the different parts of the ISO 8000-1x0 family, and (2) to provide an interface for an easier communication between the applications and the solution. These mechanisms, adequately set up within the applications requesting data, make easier the implementation of the demand of the various services from servers (data provider applications). For the sake of the widest usability and taking care of heterogeneity, we decided to focus the scope of the implementation to those applications exchanging data by means of web services. We called I8K to the set of services; and ICS-API to the API. From our approach of the problem, two applications (one of them is a data provider) willing to exchange master data with adequate levels of quality by means of web services, should adapt their regular way of working from the schema shown in Figure 1.A to the depicted in Figure 1.B.

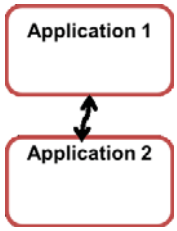


Figure 1.A. Applications exchanging data without I8K

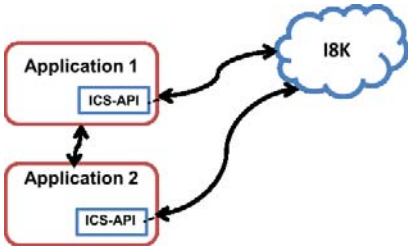


Figure 1.B. Application exchanging data with I8K

I8K has been designed as a service architecture composed of different and independent, although connected modules aimed to satisfy the different requirements imposed by the various parts of the standards. ICS-API provides classes and methods to developers so that they can build applications to request the different functionalities to I8K, like coding and decoding master data, asking for data with specific requirements of data quality, and so on.

This paper describes these two components with their corresponding elements that implement the requirements imposed by the family of standards, and finally introduces and example of use of the proposed solution. The elements that have been developed are a Data Dictionary, a communication protocol between modules and applications in which some types of messages have been identified, the schema of the master data messages implemented by means of XSDs [8], and some concerns regarding to the measuring and certification schemas. The example presents how two applications in the domain of travel agency (an application named TripPlanner requesting data and a data provider for flights) have been adapting to use the ICS-API in order to consume the services provided by I8K.

The remainder of the paper is organized as follows: In Section 2, some foundations about data quality and a description of the different parts of the standards are remembered to better understand the remainder of the paper. Section 3 introduces I8K and ICS-API and the elements that have been developed. Section 4 is aimed at explaining the way of working of the solution by means of the example, in which the use of the protocol communication governing the exchange of data is shown. And finally, conclusions are drawn and future work proposed.

2. FOUNDATIONS OF THE WORK

With the aim of adequately ground our proposal, it is necessary to analyze various arenas. On the one hand, it is necessary to analyze the most important aspects of Data Quality Management and Master Data Management. And on the other hand, the family of standards ISO/IEC 8000-1x0 is briefly described.

2.1. Master Data Management and Data Quality

The most and widest accepted definition of data quality is grounded on the vision *fitness for use* [9]. Another interesting definition of data quality is *data that meet requirements* [3]. Organizations, throughout his life-cycle, collect a lot of data according to the common business objects. These data are adapted to make them operable to applications and/or organizations [2]. Key objects supporting organizational knowledge are called Master Data [10]. Master Data are critical entities, relationships and attributes for organizations, and at the same time, also a key for the BP and application systems [2]. The Master Data Management, henceforth referred as MDM, provides solutions to problems in the data exchanged between organizations, e.g. different names for the same concepts or on the contrary, different concepts having the same name [2]. Loshin in [2] defines MDM as a collection of best practices about data management oriented to the integration of data. These practices are integrated into the business applications, information management, and tools for the management of data that implement procedures, services and infrastructure to support the capture, integration and sharing of the Master Data. The MDM has associated benefits such as increasing the level of data quality [11]. An organization can undertake actions of evaluation and improvement of data quality without taking into account mechanisms for the MDM. However, the opposite situation is not possible, since the MDM implies data quality [12].

2.2. ISO/IEC 8000-1X0:2009 Family of Standards

The main requirement of ISO/IEC 8000-1X0:2009 family of standards is to use a pre-established format for the Master Data Messages exchanged in the communications between applications. In addition to this requirement, information about the level of data quality level contained in the Master Data Message should be also added to the message.

The specific goals of each part of ISO/IEC 8000-1X0:2009 family of standards are:

- ISO 8000-100:2009 [13] describes generic aspects of Master Data to be managed.
- ISO 8000-102:2009 [14] describes the vocabulary related to the quality of master data used in the different parts of the standard.
- ISO 8000-110:2009 [4] establishes the rules that must be used for the coding of master data messages (syntactic aspects, semantic codification and requirements).
- ISO 8000-120:2009 [5] establishes the way in which the information about the life cycle of the data is added to the message to describe its/their provenance (data provenance).
- ISO 8000-130:2009 [6] establishes how to add the information regarding to the certification of accuracy of data.
- ISO 8000-140:2009 [7] establishes how to add the information regarding to the certification of completeness of data.

3. I8K AND ICS-API: AN IMPLEMENTATION OF ISO 8000-1X0

As said it was explained in the introductory section, the main contributions of this paper are I8K and ICS-API as an implementation of the ISO 8000-1x0 to support the exchange of master data by means of messages. This section introduces description of both of them and also provides details about the different elements required to use them.

3.1 I8K: A Service Architecture to support the exchange of MDM with certification of data quality levels

3.1.1. Service Architecture

The Service Architecture I8K is composed of the modules represented in Figure 2. These modules implement functionalities to meet the requirements of the different parts of ISO/IEC 8000-100:2009 family of standards. The messages exchanged from the applications to I8K and between the different modules are performed by using Web Services. The Web services have been developed using Java and Axis 2. The WSDL [15] containing the interface to access the services provided by I8K can be accessed at http://alarcosj.esi.uclm.es/i8k/.

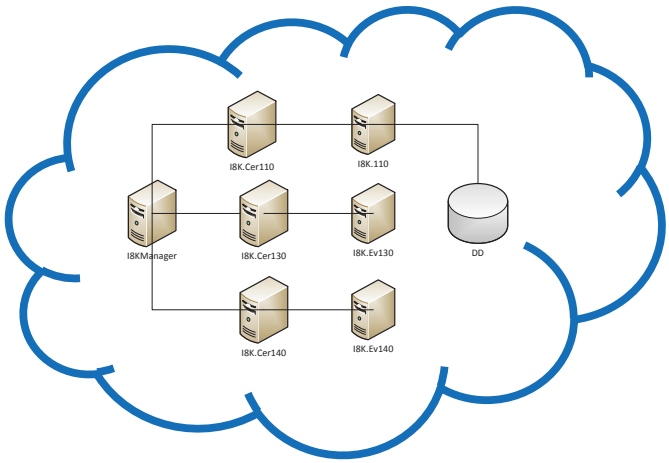


Figure 2. I8K, the service architecture

The specific responsibility of each module is described below:

- **I8KManager** serves as facade of I8K. It is responsible for receiving and processing requests that are made to the I8K from the applications. These requests include the necessity to encode, decode and/or certificate the level of quality of the data.
- **I8K.Cer110** is responsible for the management and maintenance of the data dictionary that requires the part ISO 8000-110.The master data will be encoded and decoded according to the terms representing the organizational knowledge. Each domain has their own set of terms representing the specific master data of the corresponding business.
- **I8K110** is in charge of encoding and decoding the Master Data Messages, according to the specific syntactic requirements of the standard part 110.
- **I8K.Cer130** is responsible for adding the information about certification of the level of Accuracy of the data contained in the message, according to the requirements specified in the part 130 of the standard. This information about the level of accuracy is obtained from the corresponding evaluation. This evaluation can be done by an external module or by the I8K.Ev130 module (explained below).
- **I8K.Cer140** is responsible for adding the information of certification about the level of Completeness of the master data contained in the message, according to the requirements specified in the part 140 of the standard. This information about the level of completeness is obtained from the corresponding evaluation. This evaluation can be done by an external module or by the I8K.Ev140 module (explained below).
- **I8K.Ev130** is responsible for evaluating the Accuracy of master data contained in the master data message, according to the data requirements for accuracy defined. In this implementation, these requirements are included as part of the master data message.
- **I8K.Ev140** is responsible for evaluating the Completeness of master data contained in the master data message according to some data requirements for completeness that analogously to the accuracy are included as part of the master data message

In addition, all modules add information about the life-cycle of the data (data provenance) as a requirement shown in ISO/IEC 8000-120:2009. In this way, the master data messages, which are later exchanged by applications, meet the requirement of the standard ISO/IEC 8000-100:2009. In any case, it is interesting to emphasize that for the sake of simplicity, any application requiring the functionality of I8K will communicate only with I8KManager by means of the Master Data Messages. I8KManager will analyse the information provided within the master data message, and depending on the type of message, it will redirect to the specific module the requested service. Following, a description of the format of the Master Data Messages as used in our implementation is shown.

3.1.2. Format of the Master Data Messages.

In order to enable an efficient communication between the applications and I8K, some information has to be added to the master data messages that are exchanged. This information is structured in several parts that are described below. This structure has been implemented by means of XSD. Some examples of the various parts of the master data messages (part of the XML files generated by using the previously mentioned XSD) will be shown in Section 4 with the illustrating example:

- **Header** (represented by the element ***<head>***), which contains the following information:
  - ***type-message*:** It allows the specification of the type of master data message (see Table 1 in Section 3.1.4 for the description of the types of master data messages).
  - ***syntax**:* It is used to identify the syntax information that should meets the master data message. The elaboration of the terms containing the corresponding syntax should be jointly agreed by the most important actors for each domain. In addition, some specific standards, as ISO 22745, can be used to depict this syntax as a good practice of Data Cataloguing. The Data Dictionary

contains the terms associated to each syntax that is required. At this point, it is important to highlight that the use of the services provided by I8K is generic, being necessary only to set up the corresponding terms in the Data Dictionary.

- ***cert130:*** It allows the specification of the requirement of certifying the level of data quality regarding to accuracy (*certificated130*). It also permits to provide information about the requirement of the minimum level required by the application requesting data (*requiredlevelthreshold130*), and also about the level of quality provided by the supplier (*certifiedmeasuredlevel130*). The information contained in this element, and in the following one, is necessary as part of the operative of the communication protocol, as it will be shown in subsection 3.1.4.
- ***cert140*:** This element is used to specify whether or not it is required to certify (*certificated140*) the data according to the completeness, the required minimum level (*requiredlevelthreshold140*) and the level provided by the supplier (*certifiedmeasuredlevel140*).

- **Body**, which contains the data to be exchanged between applications (attributes being queried), or the data that is exchanged (returned values satisfying the queries). In order to adequately structure the corresponding fields, some elements are nested in this part of the master data message. It will be represented by the following element <**Data>.** This part is composed of the terms (master data) that the message contains, with their respective values accordingly to the requirement of the ISO 8000-110:2009 of codifying the data as the pair (property, value). It contains the following attributes:
  - ***property-value property-ref,*** used to specify the master data
  - ***controlled-value value-ref,*** used to specify the value of the master data.
- **Data Quality Rules**, which allow the specification of the data quality rules so that I8K can perform operations to assessment and certification of quality levels. The rules used to assess the accuracy dimensions described by using the attribute *cer130*, whereas the rules to assess the completeness are depicted by means of *cer140*. To express the rules for *cert130*, the following attributes are made available to developers as result of our research:
  - ***term***: it is the master data being object of the data quality rule.
  - ***pattern***: a pattern (regular expression) that must be met by the value provided by the provider of the master data.
  - ***source*:** source of information where the value provided by the data provider can check its correctness.
  - ***required***: it indicates whether it is obligatory that the term has value.

  To express the rules for *cert140,* the following attributes are provided to developers:
  - ***term***: it is the master data concept being object of the data quality rule.
  - ***dqproperty***: it is used to mark as not null (required existence) the master data identified by the term.
- **Information provenance** (represented by the element <***provenance***>)
  This element contains the information about the life cycle of the term when it has been being exchanged, used and/or updated by different organizations. The following attributes are provided to designers to add this information:
  - ***date***: it is used to set the time and date in which the master data message is received.
  - ***event-type***: it indicates the type of action performed on the message (it takes the value of encode or decode).
  - ***organization-ref***: it represents the organization that performs the *event-type* on the master data message.
  - ***person-ref***: application that performs the operation on the message master data.
  - ***person-destination***: organization/application to which/whom the message is forwarded.
- **Accuracy Certification Information**, represented by <***accuracy-event***>. This part describes the certification information of the accuracy of the data contained in the message of master data.

- **Completeness Certification Information**, represented by ***<completeness-event>***. This part describes the certification information of the completeness of the data contained in the message of master data

  Both *Accuracy-event* and *completeness-event* contain the same information; *date*, *organization-ref* and *event-type* (it takes the value *certify130* or *certify140* respectively), where their values specifies also the provenance as defined before.

3.1.3 Data Dictionary and Associated Data Model

ISO/IEC 8000-110:2009 requires the existence of a **Data Dictionary (DD)** where all the terms corresponding to master data are stored. These master data enables the architecture of services I8K to carry out the semantic encoding and decoding of master data contained in the different master data message; these messages are exchanged between applications. The data model, for the definition of the DD consists of the following elements:

- **Term**: this field permits to specify the terms included in the vocabulary, i.e. their value in text. For example, *FROM_LOCATION*.
- **State**: it represents whether a term is active or not within the DD.
- **Language**: it specifies the language in which the term is in the DD. The same term can be stored in different languages.
- **Organization**: it is the information about the organization which has introduced the term in the DD.
- **Definition**: definition of the term. If the term is stored in multiple languages, it will have the corresponding definitions in each language.
- **Identifier**: this field contains the value of the encoded term. Therefore, when a message of master data is encoded, the term value is replaced by the value of this field.
- **Organization name**: corresponding to the name of the organization that stores the term.

As well as storing the information relating to each master data (term), it is also necessary to store the information of the formal syntax that meet the master data message. In Section 4, an example of a portion of a specific data dictionary will be provided.

3.1.4. Communication Protocol

In order to govern the functionalities of I8K, a communication protocol has been created. This protocol regulates the exchange of master data messages between applications (whose development has been largely simplified by using ICS-API) and I8K. Depending on the service that is going to be requested to I8K, developers will have to set up a specific type of master data message using ICS-API facilities. Later, when the master data message raises the I8K, the I8KManager will process the type of message and will run the corresponding set of actions to satisfy the request. Table 1 lists the message types of master data that are exchanged between applications and the architecture of services I8K. The various types of master data messages are represented in Figure 3. The number associate to each edge indicates the order in which the master data message occur. The Figure 3 also represents the direction of the different type of master data messages between the application requesting data and the application providing data.

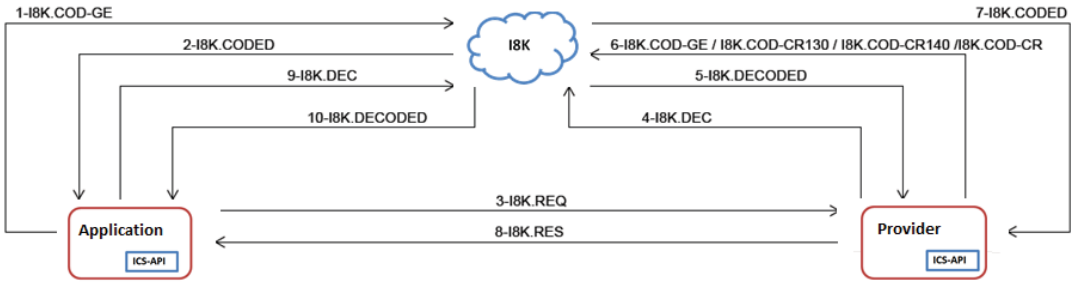


Figure 3. Protocol for the exchange of data quality levels

Type	Description
I8K.COD-GE	An application needs to encode master data to make a request to a service
I8K.CODED	I8K has codified a master data message, and the content is forwarded back to the calling application.
I8K.DEC	An application needs to decode a master data message received to understand the content
I8K.DECODED	I8K has decoded a message master data, and the content is forwarded back to the calling application.
I8K.COD-CR130	An application needs to encode the message and assess and also certify the *accuracy* of the data.
I8K.COD-CR140	An application needs to encode the message and also certify *completeness*
I8K.COD-CR	An application needs to encode the message and assess and certify the data master message, according to the quality levels of *accuracy* and *completeness*

Table 1. Types of Master Data Messages exchanged between the applications focus of our approach and I8K

The types of master data messages that are exchanged between applications are shown in Table 2.

Type	Description
I8K.REQ	An application sends a data request message to a data provider
I8K.RES	A data provider sends back a response message with the data which has requested an application

Table 2. Types of Master Data Messages exchanged between applications

3.2. ICS-API: Interface Communication Service - Application Programming Interface

ICS-API is an Application Programming Interface that the third-party applications (applications and providers) should use in order to simplify their communications with the I8K services architecture. It is implemented by using Java™ language. ICS-API offers applications the necessary primitives to:

- Set up information about the organization who owns the applications exchanging data.

- Set up the type of Master Data Message.
- Set up the requirements of certification desired.
- Perform communication with I8K to encode messages.
- Perform communication with I8K to decode messages.
- Manage the request of quality level certification of the data contained in a Master Data Message.

ICS-API has been made available to developers as an Eclipse plugin through the Eclipse Marketplace. Anyway, it is planned as future work to make the architecture available to any other IDE platforms.

The use of the various primitives of ICS-API will be illustrated by means of the example presented in Section 4.

4. WORKING EXAMPLE

In order to illustrate the use of I8K and ICS-API, an example of application is presented. In this example, an application, called TripPlanner, requests some data about flights, car renting and hotels from different providers to select the cheapest trip between two cities including flights, car rental and hotel staying. In order to organize the trip, some master data messages are exchanged between TripPlanner, and the corresponding providers. More specifically, we are only going to illustrate the exchange of information between TripPlanner and FlightCIA (flight data provider). See Figure 4 for further understanding on the communication channels.

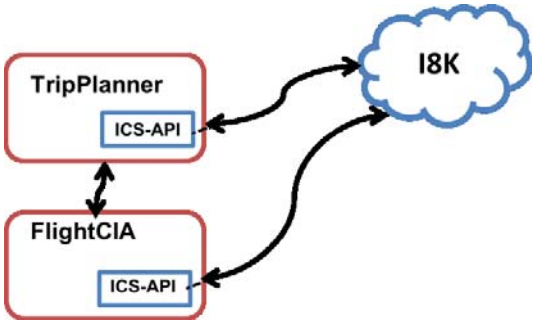


Figure 4. Communication between the different applications using the services provided by I8K

Please, notice that without the need of using ISO 8000-1x0, TripPlanner developers would simply invoke some web services aimed at satisfying the request. Having to use ISO 8000-1x0, and with the possibility of using I8K, this request has to process some types of master data message containing coded data. This extra - processing will increase undoubtedly the complexity of the process, although the main advantage is that, at the end, final users of TripPlanner will have available only data with an adequate level of quality. In the following, it is necessary to take into account that the master data message is represented by means of an XML string having the elements described in Section 3.1.2, and that this XML file is being built by means of execution of the corresponding primitives of ICS-API that can interact directly on the string or that can retrieve some information from I8K.

Very briefly, the pre-requisites that must be satisfied are the following ones:

1. I8K Service Architecture must be running and prepared to listen requests from TripPlanner and FlightCIA, process any kind of messages, and send them back the results.
2. Business and IT people from the organizations running TripPlanner and FlightCIA have to raise an agreement on the syntax to be used (set of terms representing the master data). It may occur that there already exists an international standard especially addressed at the domain (like HL7 for eHealth) or that both organizations should develop one ad hoc, which desirably can be standard in the future.

3. FlightCIA Developers should use ICS-API to rewrite and overload the functionalities aimed to process the data requests with the new features imposed by ISO 8000-1x0, which implies the usage of the services provided by I8K
4. TripPlanner Developers should use ICS-API to rewrite and overload the functionalities aimed to generate the data request with the new features imposed by ISO 8000-1x0, what implies the usage of the services provided by I8K. Probably, some information about the level of quality of data would be contained in the master data message which is returned from FlightCIA to TripPlanner. In this case, TripPlanner developers can use ICS-API to retrieve such information and on it, check the fitness for use of the data provided by means of some conditions that can help to make a decision on whether or not to use the data.

Following with the details. For the pre-requisite 1, and as previously said, I8K is running and services can be accessed at http://alarcosj.esi.uclm.es/i8k/.
For the pre-requisite 2, we can consider that Business and IT people of the two organizations have already raised an agreement on the syntax to be used. These terms should be stored in the Data Dictionary prior to the execution. Table 3 gathers, as an example, some of these terms for the domain of travels. It is interesting to highlight that the identifiers (third column) have been coded according to the standards ISO 6523 and ISO 11179-6. Let us suppose that this is part of the version 1 of a syntax called "*Classical*".

Term	Description	Identifier
SET_RETURN_DAYS	Set of possible return dates.	0102#00000008#1
FROM_LOCATION	The city where the user prefers to depart.	0102#00000001#1
TO_LOCATION	The city where the user prefers to go.	0102#00000002#1
FROM_LOCATION_FLIGHT	City from flight departs.	0102#00000014#1
FROM_AIRPORT	Airport from which the flight departs.	0102#00000016#1
FLIGHT_NUMBER	Flight number.	0102#00000022#1
CARRIER	Airline which operates the flight.	0102#00000023#1
FLIGHT_PRICE	Flight price.	0102#00000024#1

Table 3. Terms corresponding to the domain of travel, terms which have to be stored in the Data Dictionary

For achieving the pre-requisites 3 and 4, it is necessary that TripPlanner and FlightCIA developers download and set up the Eclipse plugin ICS-API to their corresponding projects.

On both sides, developers will use the ICS-API class MDQManager's primitives to set up some details necessary to communication between TripPlanner and I8K, and data providers and I8K. This includes some concerns such as the name of the application, the name of the organization, the name and version of the syntax, the access point of I8K… As an example, developers could use a code like the following one to specify who makes the request, in the example is the TripPlanner application, and for whom is the request, in this example, for the FlightCIA (flight provider):

```
MDQManager manAP=new MDQManager();
manAP.configureOrganization("TripPlanner", OrganizationType.AP);
manAP.setDestination("FlightCIA");
...
```

On TripPlanner's side, developers must prepare the master data message that will contain the request of data about flights from FlightCIA. This is, according to Table 2, an I8K.REQ message type, in which in addition to data itself, some data quality requirements are fixed. Supposing that the specific data quality requirements do not need to certify the level of accuracy (ISO 8000-130) for the data from FlightCIA to TripPlanner, and the level of certified completeness (ISO 8000-140) must be at least 50% for this data. By using the following code, ICS-API will be preparing a master data message whose header is represented in Figure 5.

manAP.configureCertification(0, 50, false, true);

```
<head>
        <type-message type="I8K.REQ"/>
        <syntax syntax_version="1.0" syntax_name="Classical" syntax_id="1"/>
        <cert130 requiredlevelthreshold130="0" certificated130="false"/>
        <cert140 requiredlevelthreshold140="50" certificated140="true"/>
</head>
```

Figure 5. Header of the I8K.REQ message for our illustrative example

Corresponding the first parameter to the minimum threshold of acceptable level of accuracy for the data being requested; the second parameter for the minimum threshold of acceptable level of completeness for the data being requested, and finally, the third and the fourth should be set to *true* if certification of the level of quality are required for ISO 8000-130 and ISO 8000-140 respectively.
But previously to generate this specific I8K.REQ message, TripPlanner, in order to be ISO 8000-1x0 compliant, and once decided which master data is required to be queried to FlightCIA will have to ask to I8K to encode the data. For instance, to query which flights from Seville to London, FlightCIA can offer to TripPlanner, developers can use this fragment of code using ICS-API primitives:

manAP.addTermAndValue("FROM_LOCATION", "Seville");
manAP.addTermAndValue("TO_LOCATION", "London");
...
message=manAP.encodeAndCertificated();

This request is done to I8K by means of a master data message of type I8K.COD-GE, which will return a master data message of type I8K.CODED containing in its body section the fragment XML represented in Figure 6. Later, TripPlanner developers will introduce this data into the I8K.REQ message to be sent to FlighCIA by using the corresponding ICS-API primitivesThe property-value node stores the name or identifier of the master data (terms) and controlled-value node stores the value taken.

```
<data>
        <property-value  property-ref="01-02#00-000008#1">
                <controlled-value  value-ref="2013-06-15"/>
                <controlled-value value-ref="2013-06-16"/>
        </property-value>
                        …
        <property-value  property-ref="01-02#00-000001#1">
                <controlled-value  value-ref= “Seville"/>
        </property-value>
</data>
```

Figure 6. Fragment corresponding to coded data of the master data message of type I8K.CODED from I8K to TripPlanner

As part of the I8K.REQ message, TripPlanner developers will have to coherently describe the data quality requirements that FlightCIA must satisfy. This can be understood as a kind of Service Level Agreement. To do so, ICS-API provides developers with primitives that enable them to build the section <data-quality-rules> of the I8K.REQ message. For instance, to specify the rules for completeness, developers can mark the data which cannot be null in the I8K.RES message that FlightCIA will send back to TripPlanner satisfying the data request by means of the following code:

manAP.addTermRequired("FLIGH_PRICE", true);
manAP.addTermRequired("CARRIER", true);

Or to deal with accuracy, developers can specify a regular expression that data should satisfy or specify an authoritative source that can validate the value of the data. In the example, regarding rules to assess the level of **accuracy** of master data shows that the value of the term 01-02#00-000013#1 has to fulfill the specified pattern. As for the rules to evaluate the degree of completeness it is noted that the terms specified values cannot be empty. To do so, developers can use code like this:

manAP.addTermPattern(CHECK_IN", "^((19|20)dd)-(0?[1-9]|1[012])-(0?[1-9]\|[12][0-9]|3[01])$", true);
manAP.addTermSource("DESTINATION", "http://wordnet.princeton.edu/", true);

As a result, the <data-quality-rules> will look like as shown in Figure 7. In addition to this information, the track of the life cycle is stored as part of the master data message within the section <provenance>. This information is added to the master data message when this master data message flows between modules. Figure 8 shows the section <provenance> of the master data message that TripPlanner will send to FlightCIA.

```
<data-quality-rules>
        <cert130>
                <set term="01-02#00-000013#1" pattern="((19|20)dd)-(0?[1-
                9]|1[012])-(0?[1-9]|[12][0-9]|3[01])$" required="true"/>
                <set term="01-02#00-000025#1"
                source="http://wordnet.princeton.edu/" required=true
        </cert130>
        <cert140>
                <set term="01-02#00-000024#1" dqproperty="required"/>
                <set term="01-02#00-000022#1" dqproperty="required"/>
        </cert140>
</data-quality-rules>
```

Figure 7. Fragment corresponding to <data-quality-rules> corresponding to the master data message of type I8K.REQ that will be sent from TripPlanner to FlightCIA

```
<provenance>
        <provenance-event person-destination="FlightCIA" person-ref="TravelAgency" organization-
        ref="I8K" event-type="encode" date="2013-05-24T18:26:57.891+02:00"/>
        <provenance-event person-ref="I8KManager" organization-ref="I8K" event-type="encode"
        date="2013-05-24T18:26:50.338+02:00"/>
        <provenance-event person-ref="AgI8KCer110" organization-ref="I8K" event-type="encode"
        date="2013-05-24T18:26:50.354+02:00"/>
        <provenance-event person-ref="AgI8K110" organization-ref="I8K" event-type="encode"
        date="2013-05-24T18:26:50.541+02:00"/>
        <provenance-event person-ref="AgI8KCer110" organization-ref="I8K" event-type="encode"
        date="2013-05-24T18:26:50.557+02:00"/>
        <provenance-event person-ref="I8KManager" organization-ref="I8K" event-type="encode"
        date="2013-05-24T18:26:50.572+02:00"/>
</provenance>
```

Figure 8. Information of lifecycle of data of a master data message

Once the master data message type I8K.REQ is ready, developers must invoke the web service of FlightCIA, which is in charge of satisfying the request. Basically, the functionality is the same, since FlighCIA should already have a web service to provide data about flights. The main difference is that this

former service has to be now encapsulated in a new one which has to take into account the specific details of our ISO 8000 implementation. This difference is that now the exchanged objects are not only the parameters of the flights, but the master data message. This implies that the processing of the request will not only produce the data satisfying the request, but the data which in addition of satisfying the request does also satisfied the data quality requirements that are described as part of the I8K.REQ message. Moreover, if certification of data quality levels is required, I8K will act as authoritative source to add the corresponding information about this or these level(s) of certification (by now, the only ones that are covered by the standard: accuracy and completeness).
Consequently, at the FlightCIA side, developers must use ICS-API functionalities to build master data messages of type I8K.DEC that will be sent to I8K to decode the terms of the data request. As a result, I8K will produce a master data message of type I8K.DECODED with the data to be queried.
Once received the message, FlightCIA will properly execute the query (flights from Seville to London on specified dates). Without the need of ISO 8000-1x0, FlightCIA could send right now the answer to TripPlanner, and the process will finish: TripPlanner will have the data that satisfies the request (e.g. Flight IB0123 departing from Seville on 15-06-2013 and arriving in London on 15-06-2013), but without any kind of warranties that the returned data really satisfies the data quality requirements. To add this warranty, developers, using ICS-API primitives, will invoke again I8K services to code the terms, to assess the level of quality of the data that satisfy the data request, and to add some information (like the corresponding certification) about the fitness for use of the data. This is to be done by means of the master data messages I8K.COD-GE (only encoding but no certification is required), I8K.COD-CR130 (encoding and certification of ISO 8000-130), I8K.COD-140 (encoding and certification of ISO 8000-140) or I8K.COD-CR (encoding and certification of both ISO 8000-130 and ISO 8000-140). Depending on the type of required certification, I8K will add any or both of them to the fragment shown in Figures 9 and 10 (as required previously – see Figure 5) to the mater data message of type I8K.CODED that is returned to FlightCIA.

```
<completeness-event xsi:nil="true"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"/>
```

Figure 9. Information of certification about the level of completeness of master data

```
<accuracy-event date="2013-06-03T11:27:27.007+02:00"
organization-ref="I8K" event-type="certify130">
        100%
</accuracy-event>
```

Figure 10. Information of certification about the level of accuracy of master data

With all of this information, FlightCIA can build the master data message of type I8K.RES that will be sent back to TripPlanner to satisfy the data request.
Once TripPlanner has the message, the message has to be decoded so that data can be used. This implies a new exchange of master data message of types I8K.DEC and I8K.DECODED. To process the information regarding to certification, ICS-API provides primitives that can be used as follow:

```
if (manAP.getLevel130provided() and manAP.getLevel140provided())
        ShowDataToUser();
else
        throw new DataQualityException();
```

In the example that we use to build the tutorial of ICS-API, we ran TripPlanner against a flight provider to find the cheapest flights from Seville to London, in two scenarios, without asking data having certified levels of data quality and asking data with the level of completeness certified. We obtain that the cheapest flight in the first scenario (without using I8K and ICS-API) had a cost of 290 €, whereas in the second

case, we obtained a flight with a cost of 320.40 €. In our tests, both results were offered to TripPlanner users, rejected the first offer, because trusting data was more important that having cheaper flights. This led us to check that how companies can lost customers due to inadequate levels of quality, because, in spite that data representing the flights could have not adequate level of quality, however, the flights per se would be correct, and the user could become totally satisfied with the flight, but users do not want to risk their money. Anyway, it is also important to highlight the cost in terms of computational or communication performance loses that using this solution can bring.

5. CONCLUSSIONS AND FUTURE WORKS

In this paper, we introduce an implementation of ISO 8000-1x0 family, which covering requirements specified by the cited part of the standards, works under the assumption that master data is exchanged by means of Web Services between applications requiring to deal data with some data quality requirements. The implementation we have introduced has two main elements, on a hand, we have described I8K, a service architecture based on Axis 2 Java Web Services that has several modules that perform the operations of encoding and decoding (part 110). The architecture also permits to add data provenance information (part 120), and assessing and if demanded certifying the level of quality of data of the only two data quality dimensions covered by the standard in parts 130 (accuracy) and 140 (completeness). I8K has also a database containing the master data that application would need for an effective communication. Along with this service architecture, we have also developed an API, called ICS-API, which provides developers with primitives that enable the consumption of the services provided by I8K.

As the main advantages of the implementation, we can say that I8K can be used for the exchange of data in any other domain (the only requirement is to introduce new syntax – terms of mater data- into the Data Dictionary), it is scalable and extensible with new data quality dimensions since the communication protocol and the data model that we have developed as part of our research will allow these new features. As the main disadvantages we can cite the need of rewrite the applications with ICS-API that organizations wanted to exchange data have to do to satisfy the requirement of ISO 8000-1x0, the number of calls to I8K web services that applications has to do and the effects on the performance of the applications, and some concerns regarding to the security. Also, as a disadvantage, we can cite the low number of data quality dimensions that can be assessed and certified, but this is a matter of the parts of ISO 8000-1x0.

As part of our future work, we want to: 1) add more data quality dimensions to the assessed and certified by I8K, 2) we want to refine I8K by adding some features regarding to security, 3) we also want to deal with performance concerns that will imply changes to I8K and ICS-API and 4) conduct pilot project with organizations from different domains alone and simultaneously that have different volumes of traffic data so that we could tune up adequately to optimize I8K.

ACKNOWLEDGEMENT

This work has been partially supported by the research projects GEODAS (TIN2012-37493- C03-01), funded by the Ministry of Economy and Competitiveness; MAGO (TIN2009-13718-C02-01), Ministry of Science and Technology of Spain; and TDiaCO-BPMS (TIN2009-13714), funded by the European Regional Development Fund (ERDF / FEDER).

REFERENCES

1. Wang, R., *A Product Perspective on Total Data Quality Management.* Communications of the ACM, 1998. **41**(2): p. 58-65.
2. Loshin, D., *Master Data Management*2009, Burlington, MA, USA: Morgan Kaufmann.
3. Benson, P. and M. Hildebrand, *Managing Blind: A Data Quality and Data Governance Vade Mecum*2012, Bethlehem (Pensylvania): ECCMA.
4. ISO/IEC. *ISO/IEC 8000-110: Master Data: Exchange of characteristic data: Syntax, semantic encoding, and conformance to data specification*. 2009.
5. ISO/IEC. *ISO/DIS 8000-120: Master Data: Exchange of characteristic data: Provenance*. 2009.
6. ISO/IEC. *ISO/DIS 8000-130: Master Data: Exchange of characteristic data: Accuracy*. 2009.
7. ISO/IEC. *ISO/DIS 8000-140: Master Data: Exchange of characteristic data: Completeness*. 2009.
8. Fallside, D. and P. Walmsley, *XML Schema Part 0: Primer - Second Edition*, 2004, World Wide Web Consortium: http://www.w3.org/TR/xmlschema-0/.
9. Batini, C. and M. Scannapieca, *Data Quality: Concepts, Methodologies and Techniques*2006: Springer-Verlag Berlin Heidelberg.
10. Otto, B. and A. Schmidt, *Enterprise Master Data Architecture: Design Decision and Options*. 2010.
11. Otto, B. and A. Reichert, *Organizing master data management: findings from an expert survey. Proceedings of the 2010 ACM Symposium on Applied Computing.* 2010.
12. Hayler, A., *Data quality essential to master data management.* 2012.
13. ISO/IEC. *ISO/DIS 8000-100: Master Data: Exchange of characteristic data: Overview*. 2009.
14. ISO/IEC. *ISO/IEC 8000-102: Master Data: Exchange of characteristic data: Vocabulary*. 2009.
15. Christensen, E., et al., *Web Services Description Language (WSDL) 1.1*, 2001, World Wide Web Consortium: http://www.w3.org/TR/wsdl.